Satosi Watanabe

Information Theoretical Analysis of Multivariate Correlation

Abstract: A set λ of stochastic variables, $y_1, y_2, \ldots, y_n$, is grouped into subsets, $\mu_1, \mu_2, \ldots, \mu_k$. The correlation existing in λ with respect to the μ's is adequately expressed by $C = \sum_{i=1}^{k} S(\mu_i) - S(\lambda) \geq 0$, where $S(\nu)$ is the entropy function defined with reference to the variables y in subset ν. For a given λ, C becomes maximum when each μ_i consists of only one variable, $(n=k)$. The value C is then called *the total correlation in* λ, $C_{tot}(\lambda)$. The present paper gives various theorems, according to which $C_{tot}(\lambda)$ can be decomposed in terms of the partial correlations existing in subsets of λ, and of quantities derivable therefrom. The information-theoretical meaning of each decomposition is carefully explained. As illustrations, two problems are discussed at the end of the paper: (1) redundancy in geometrical figures in pattern recognition, and (2) randomization effect of shuffling cards marked "zero" or "one."

1. Introduction

Since the time of Boltzmann, physicists have repeatedly encountered quantities of the type

$$S = -\sum_i p_i \log p_i \,, \tag{1.1}$$

either as an expression of thermodynamic entropy or as a quantity closely related to it.[1] Boltzmann seems to have noticed that thermodynamic entropy has something to do with orderliness and disorderliness of elements, and Szilard[2] in 1929 pointed out that a decrease of thermodynamic entropy is accompanied by the acquisition of information. In 1932 von Neumann,[3] using a quantity (1.1) as a model of thermodynamic entropy, demonstrated the *irreversibility of observation*, in which information about the state of a system becomes statistically less accurate. Less well known, however, is a paper[4] on atomic nuclei which appeared in 1939, using the quantity (1.1) in a practical way for the first time as an instrument deliberately contrived to measure the uncertainty of the state of a particle, and making clear that *it is different from thermodynamic entropy*. This paper may be considered, in a way, as one of the earliest events which led to the independence of information theory from thermodynamics.

In any event the 1939 paper is pertinent to our present study since it clearly stated that the quantity (1.1), on the one hand, measures the uncertainty of our knowledge about the state but, on the other, it also measures the strength of the correlation (beyond the average interaction) among constituent particles. We shall presently sketch the argument which was then used to justify this statement and show how it ties in with the present theory. It is noteworthy that the relation between information and correlation was already demonstrated at that early stage of development of information theory.

In the field of communication theory, as early as 1928, Hartley expressed the idea that the quantity of information might reasonably be defined as the logarithm of the number of independent signals.[5] This is obviously a special case of (1.1) with equal probabilities, but a formal information theory in communication based on the quantity (1.1) was developed much later by Shannon[6] and others, who adopted the statistical point of view of communication emphasized by Wiener,[7] Kolmogoroff,[8] and others. In communication theory, the relation between information and correlation was rediscovered in connection with so-called *redundancy*. After the nature of redundancy in communication became well understood, the

same notion was again applied to the problem of correlation existing in a collection of stochastic variables. The importance of this aspect of information theory in the problem of organization was emphasized with prophetic zeal by J. Rothstein.[9] Several other authors have also proposed and tried to use information quantities in analysis of multivariate correlation, among them W. J. McGill,[10] W. R. Gardner,[11] and S. K. Mitra.[12]

With regard to the correlation existing in a stationary time sequence of stochastic variables, a convenient quantity $W^{(r)} \geq 0$, called *correlation index* of range r, was introduced,[13] which measures the strength of correlation of length r in excess of the correlation of length $r-1$. It was shown that the total correlation can be expanded in terms of $W^{(r)}$ with $r=2, 3, 4 \ldots$. This applies to an infinitely long sequence, on which a segment of a given length has the same probability of being in a given state, no matter where this segment is taken.

Another simple case is a set of a finite number of stochastic variables, whose collective probability distribution is invariant for any permutation of variables. A theory enabling one to expand the total correlation among these variables in terms of partial correlations was put forward,[14] in a certain analogy with the expansion in terms of the W's.

The present paper is a more general consideration of correlation where the symmetries among variables which we have discussed in the last two paragraphs do not necessarily exist. At the same time, the cases where these symmetries do exist will also be discussed here as special cases. Several new theorems are introduced and proven.

Let us go back to the 1939 paper in nuclear physics and try to explain the main points without using highly technical notions. Those readers who are not interested in theoretical physics are advised to proceed from this point to the next Section. In a nucleus, the nucleons are attracted to each other by two-body mutual potential (forgetting, for simplicity, the possible many-body potentials). In the Hartree approximation and the Hartree-Fock approximation, however, these mutual potentials are replaced by a smoothed-out external potential to express the average effect of the mutual potentials. In this fictitious potential, there are quantum states, ψ_i, $i=1, 2, 3, \ldots$, which can be occupied by a nucleon. In the Hartree model, each nucleon is supposed to occupy a definite quantum state. Hence, for each nucleon, p_i is zero or one. Therefore, S of (1.1) becomes zero. On the other hand, in the Hartree-Fock model, the indistinguishability of elementary particles is taken into account, and as a result each particle occupies each of the n lowest quantum states with probability $1/n$, where n is the number of nucleons in the nucleus. Thus, $p_i=1/n$ and $S=\log n$.

However, neither the Hartree model nor the Hartree-Fock model takes into account the two-body correlation which is beyond the average effect. For instance, if one wishes to find the average density (probability of presence) of a particle, the answer may be given fairly correctly by these simplified models. But if one asks the conditional probability of presence of one particle at a certain position, on the hypothesis that another particle is found at a certain other position, then these simplified models are bound to give a completely wrong answer, because the models take into account only the average force acting on a particle. Such a clustering effect of particles can take place without changing the average density. To represent such a correlative (fluctuating) interaction in theory, one naturally has to allow particles to occupy the quantum states which are not occupied in the single-particle model, such as the Hartree and Hartree-Fock models. This means, for each particle, that the probability of occupying one quantum state has to be less than $1/n$, or $p_i<1/n$, since more than n states have to be occupied. This automatically entails $S>\log n$. As we can show very easily, S can be defined in a form invariant for any unitary transformation. Hence this result, $S>\log n$, is independent of the initial approximation, i.e., of the eigenfunctions used. The expression $S-\log n$ can be considered as a measure of correlation. For that matter, $\log n$ itself can be considered as a kind of correlation imposed by the Pauli principle.

In the exact wording of the quoted paper,[4] "Let the measure of indeterminacy be called S. S measures the extent to which the wave function of the nucleus Ψ gives an *indeterminate* information regarding the state of a building block [nucleon]." Thereafter S is defined by a formula equivalent to (1.1). In another part of the paper, it is stated that $S-\log n$ "gives therefore a natural measure for the importance of the fluctuation of force field, or in other words, for the influence of exchange of energy, etc." It is also stated that "through the more exact consideration of strong interaction in pairs of neighboring particles, the degree of our knowledge of the state of a particle will become less." In any event, the double role of S, once as a measure of uncertainty of the state of a single particle and another time as a measure of correlation among particles, was the theme of the paper. In a heavier nucleus it was estimated in this paper that $S=\ln n+0.73$ (natural logarithm unit).

Now the above argument of the paper is in essential agreement with the point of view of current information theory, as elaborated in the present paper. The total correlation existing in a set of n variables, $y_1, y_2, \ldots, y_n$, is adequately represented by

$$C_{\text{tot}}=S^{(1)}(y_1)+S^{(1)}(y_2)+\ldots+S^{(1)}(y_n) - S^{(n)}(y_1, y_2, \ldots, y_n). \tag{1.2}$$

This indeed shows that correlation increases as complexity of states of individual variables increases or as complexity of states of the system of variables as a whole decreases. To be more specific, we have to reinterpret this expression in a fashion acceptable to quantum mechanics. First, each $S^{(1)}(y_k)$, $k=1, 2, \ldots, n$, corresponds to (1.1). Since all the nucleons are similar, the sum of the first n terms in (1.2) is just n times (1.1), where i labels nucleon quantum states. The last term in (1.2) can also be written in the form of (1.1), but the index i must here refer to the quantum states of a nucleus as a whole. Since we are discussing a normal (unexcited) state of a nucleus, the nu-

cleus is in a definite quantum state. Hence, $S^{(n)}(y_1, y_2, \ldots, y_n)=0$. Consequently, $C_{\text{tot}}=nS^{(1)}$. We can now see that $S^{(1)}$ can be used as a measure of correlation, C_{tot}. Then $n \log n$ is a part of correlation as imposed by the Pauli principle. In familiar cases of correlation in communication, an increase of correlation is often attained by a decrease of $S^{(n)}$, but in this example it is attained by an increase of $S^{(1)}$.

Actually, this quantum-mechanical interpretation of (1.2), although perfectly faithful to the true meaning of $S^{(1)}$ and $S^{(n)}$, does not agree with the way these quantities should be calculated in conventional statistics (to which our present paper is devoted). For instance, in the usual statistics, we always have $S^{(n)} \geq S^{(1)}$, as will be seen later in (2.20). But this is not necessarily the case in quantum statistics, which has to be used in interpreting (1.2) for our nuclear problem. In order to be rigorous, we should define the information function of an r-body system, not by a formula of the type (1.1), but by

$$S^{(r)}=-\text{Spur}\,\rho^{(r)} \log \rho^{(r)}, \tag{1.3}$$

where $\rho^{(r)}$ is the r-body density matrix and "Spur" designates the diagonal sum. A detailed discussion will be given elsewhere of the quantum-mechanical information function, (1.3), which mathematically includes the ordinary information function, (1.1), as a special case and which was historically put into practical use[4] even before the latter was introduced in communication theory.

2. Decomposition of total correlation into partial correlations

We are given a set of n stochastic variables, $y_1, y_2, \ldots, y_n$, where $y_i (i=1, 2, \ldots, n)$ can take any one of g_i different discrete values. The probability that the variables $y_1, y_2, \ldots, y_n$ take values $x_1, x_2, \ldots, x_n$, respectively, will be denoted by $p(y_1=x_1, y_2=x_2, \ldots, y_n=x_n)$, or simply, $p(x_1, x_2, \ldots, x_n)$, or still more simply $p(\lambda)$, where λ stands for

$$\lambda \equiv (x_1, x_2, \ldots, x_n). \tag{2.1}$$

The symbol λ will be used sometimes to designate also the variables $(y_1, y_2, \ldots, y_n)$ instead of their values $(x_1, x_2, \ldots, x_n)$. The probability $p(\lambda)$, naturally satisfies

$$p(\lambda) \geq 0, \tag{2.2}$$

$$(\sum_{x \epsilon \lambda})^n p(\lambda)=1, \tag{2.3}$$

where the summation symbol

$$(\sum_{x \epsilon \lambda})^n = \sum_{x_1}^{g_1} \sum_{x_2}^{g_2} \cdots \sum_{x_n}^{g_n}. \tag{2.4}$$

The set λ of n variables is now divided into two subsets μ and ν respectively containing l and m variables.

$$\mu \cup \nu=\lambda, \quad \mu \cap \nu=\phi, \quad n=l+m, \tag{2.5}$$

where ϕ is the empty set. It is *not* hereby implied that μ consists of the *first* l variables $(x_1, x_2, \ldots, x_l)$, but that it consists of a set of certain l variables taken out of $(x_1, x_2, \ldots, x_n)$. Then, we have

$$p(\mu)=(\sum_{x \epsilon \nu})^m p(\lambda), \tag{2.6}$$

where the summation is taken with respect to m x's contained in ν. Similarly,

$$p(\nu)=(\sum_{x \epsilon \mu})^l p(\lambda). \tag{2.7}$$

The functions $p(\mu)$ and $p(\nu)$ have respectively l and m arguments, and obey the non-negative condition and the normalization condition of the type (2.2) and (2.3).

The information carried by n y's in λ is

$$S(\lambda)=-(\sum_{x \epsilon \lambda})^n p(\lambda) \log p(\lambda), \tag{2.8}$$

where λ on the left should be understood as denoting the variables $(y_1, y_2, \ldots, y_n)$ while λ on the right for their values $(x_1, x_2, \ldots, x_n)$. Similarly, the information carried by μ and ν are given by

$$S(\mu)=-(\sum_{x \epsilon \mu})^l p(\mu) \log p(\mu), \tag{2.9}$$

$$S(\nu)=-(\sum_{x \epsilon \nu})^m p(\nu) \log p(\nu). \tag{2.10}$$

It is natural to consider the entropy function of an empty set of variables to be zero. It is easy to see, by virtue of Gibbs' theorem, that

$$S(\lambda) \leq S(\mu)+S(\nu), \tag{2.11}$$

where the equality holds if and only if

$$p(\lambda)=p(\mu)p(\nu) \tag{2.12}$$

for all values of $(x_1, x_2, \ldots, x_n)$. Thus (2.12) can also be rewritten as

$$p(\nu|\mu) \equiv \frac{p(\lambda)}{p(\mu)}=p(\nu), \tag{2.13}$$

$$p(\mu|\nu) \equiv \frac{p(\lambda)}{p(\nu)}=p(\mu), \tag{2.14}$$

where $p(\nu|\mu)$ is the conditional probability for ν on the assumption μ. Therefore, from (2.13) and (2.14), it can be understood that the set μ of variables and the set ν of variables are "independent" of, or "uncorrelated" with, each other. As a consequence, the loss of information (redundancy) given by

$$C(\lambda; \mu, \nu)=[S(\mu)+S(\nu)]-S(\lambda) \geq 0 \tag{2.15}$$

can be used as a measure of the strength of correlation between μ and ν. We shall sometimes refer to $C(\lambda; \mu, \nu)$ as "correlation existing in λ with respect to μ and ν."

If one observes the variables contained in μ and the variables contained in ν separately, then the information carried by μ and that carried by ν are respectively $S(\mu)$ and $S(\nu)$. But on account of correlation between the variables, the information carried simultaneously by the variables in $\lambda=\mu \cup \nu$ is $S(\lambda)$, which is less than $S(\mu)+S(\nu)$.

The difference is $C(\lambda; \mu, \nu)$. One can reinterpret this quantity in terms of the notion of "ignorance before observation," noticing the fact that the "information" is the decrease in ignorance by the observation. If one has not observed any of the variables $(x_1, x_2, \ldots, x_n)$ in λ, then one's "ignorance" about the values of the variables in μ is expressed by

$$-(\sum_{x\epsilon\mu})^l p(\mu) \log p(\mu) = S(\mu) . \tag{2.16}$$

Suppose now that one has observed the values of the variables in ν, then the ignorance about the values of the variables in μ becomes

$$-(\sum_{x\epsilon\mu})^l p(\mu|\nu) \log p(\mu|\nu) = S(\mu|\nu) . \tag{2.17}$$

The μ in $S(\mu|\nu)$ stands for the y's in μ while the ν in $S(\mu|\nu)$ stands for the x's in ν. Now these observed values of the variables in ν occur with probability $p(\nu)$. Therefore, the expected value of $S(\mu|\nu)$ is

$$\begin{aligned}(\sum_{x\epsilon\nu})^m S(\mu|\nu) p(\nu) &= -(\sum_{x\epsilon\nu})^m p(\nu)(\sum_{x\epsilon\mu})^l \frac{p(\lambda)}{p(\nu)} \log \frac{p(\lambda)}{p(\nu)} \\ &= S(\lambda) - S(\nu) . \end{aligned} \tag{2.18}$$

Without any observation, the ignorance about μ was $S(\mu)$ of (2.16). After the observation of ν, the ignorance has become on the average $S(\lambda)-S(\nu)$ of (2.18). The decrease in ignorance is the information about μ provided by the observation of ν and is given by

$$S(\mu) - [S(\lambda) - S(\nu)] = C(\lambda; \mu, \nu) . \tag{2.19}$$

If the variables in μ and the variables in ν are mutually independent, then the observation of the variables of ν would not help in any measure the prediction about the values of the variables in μ. If there is any correlation, however, this observation of ν will provide some information about the outcomes of μ. This indirect information is given by (2.19).

Relation (2.11) sets the lower limit to $C(\lambda; \mu, \nu)$, which is zero. Now Eq. (2.17) serves the purpose of setting the upper limit to it. Since $p(\mu|\nu)$ is a probability distribution for the variables in μ, $S(\mu|\nu)$ of (2.17) is non-negative. It becomes zero if and only if $p(\mu|\nu)=0$ or 1. This means that the values of the variables in μ are completely determined by the knowledge of the values represented by ν. Now, since $S(\mu|\nu)$ as well as $p(\nu)$ is non-negative, $S(\lambda)-S(\nu)$ in (2.18) is also non-negative:

$$S(\lambda) - S(\nu) \geq 0 . \tag{2.20}$$

Equality in (2.20) happens if and only if all $S(\mu|\nu)$ (for ν such that $p(\nu)\neq 0$) vanish. This means a complete dependence of μ on ν. Changing the names of μ and ν, one obtains also $S(\lambda)-S(\mu)\geq 0$, where equality holds if and only if ν is completely dependent on μ. Subtracting $[S(\lambda)-S(\nu)]$ from $S(\mu)+S(\nu)$, and using (2.20), one obtains

$$C(\lambda; \mu, \nu) \leq S(\mu) , \tag{2.21}$$

where equality occurs when μ is completely dependent on ν.

Similarly,

$$C(\lambda; \mu, \nu) \leq S(\nu) . \tag{2.22}$$

Therefore, $C(\lambda; \mu, \nu)$ is not larger than the smaller of $S(\mu)$ and $S(\nu)$. It is easy to see that if $S(\mu)$ is larger than $S(\nu)$, then it is impossible for μ to be completely dependent on ν.

Next, by dividing λ into k subsets $\mu_i (i=1, 2, \ldots, k)$, each containing l_i variables, so that

$$\begin{aligned} &\mu_1 \cup \mu_2 \cup \ldots \cup \mu_k = \lambda , \\ &\mu_i \cap \mu_j = \phi , \quad i \neq j , \\ &n = l_1 + l_2 + \ldots + l_k , \end{aligned} \tag{2.23}$$

we obtain

$$S(\lambda) \leq \sum_{i=1}^{k} S(\mu_i) . \tag{2.24}$$

Equality in (2.24) holds obviously if and only if

$$p(\lambda) = p(\mu_1) p(\mu_2) \ldots p(\mu_k) \tag{2.25}$$

for all possible values of $(x_1, x_2, \ldots, x_n)$. Thus, the quantity

$$C(\lambda; \mu_1, \mu_2, \ldots, \mu_k) = \sum_{i=1}^{k} S(\mu_i) - S(\lambda) \geq 0 \tag{2.26}$$

measures the correlation existing among subsets μ_1, μ_2, $\ldots$, and μ_k, and becomes zero when they are mutually independent, in the sense of (2.25).

The quantity $C(\lambda; \mu_1, \mu_2, \ldots, \mu_k)$ will be sometimes called the correlation existing in λ with respect to $\mu_1, \mu_2, \ldots, \mu_k$. In particular, if $(\mu_1, \mu_2, \ldots, \mu_k)$ becomes $(y_1, y_2, \ldots, y_n)$, i.e., if $l_1=l_2=\ldots l_n=1$, then $C(\lambda; \mu_1, \mu_2, \ldots \mu_k)$ will be called *total* correlation existing in λ:

$$\begin{aligned} C_{\text{tot}}(\lambda) &= C(\lambda; y_1, y_2, \ldots, y_n) \\ &= \sum_{i=1}^{n} S(y_i) - S(\lambda) . \end{aligned} \tag{2.27}$$

We shall presently see that $C_{\text{tot}}(\lambda)$ is the largest among all possible $C(\lambda; \mu_1, \mu_2, \ldots, \mu_k)$, when λ is given.

Now taking a given value μ_i, let us further subdivide it into subsets $\nu_{i,j} (j=1, 2, \ldots, k')$ such that

$$\begin{aligned} &\nu_{i,1} \cup \nu_{i,2} \cup \ldots \cup \nu_{i,k'} = \mu_i , \\ &\nu_{i,j} \cap \nu_{i,l} = \phi , \quad j \neq l . \end{aligned} \tag{2.28}$$

The correlation existing in μ_i with respect to $\nu_{i,1}, \nu_{i,2}, \ldots, \nu_{i,k'}$, which is given by

$$C(\mu_i; \nu_{i,1}, \nu_{i,2}, \ldots, \nu_{i,k'}) = \sum_{j=1}^{k'} S(\nu_{i,j}) - S(\mu_i) \geq 0 \tag{2.29}$$

will vanish if and only if

$$p(\mu_i) = p(\nu_{i,1}) p(\nu_{i,2}) \ldots p(\nu_{i,k'}) . \tag{2.30}$$

We can proceed in this fashion until finally each subset

consists of only one variable y. At each branching point, i.e., every time a subset is divided into sub-subsets, a correlation C is defined in the way indicated in (2.29). Suppose that λ is divided into the μ's, and the μ's are divided into the ν's, et cetera, until finally the subsets (κ's) are divided into the y's. Then, we have

$$C(\lambda;\mu_1,\ldots)=\sum_i S(\mu_i)-S(\lambda)$$

$$C(\mu_i;\nu_{i1},\ldots)=\sum_j S(\nu_{ij})-S(\mu_i)$$

$$\cdots\cdots\cdots\cdots\cdots$$

$$C(\kappa_l;y_{l1},\ldots)=\sum_m S(y_{lm})-S(\kappa_l)\,. \qquad (2.31)$$

Adding all the equations of this type, we obtain

$$C(\lambda;\mu_1,\ldots)+\sum_i C(\mu_i;\nu_{i1},\ldots)+\ldots.$$

$$+\sum_l C(\kappa_l;y_{l1},\ldots.)=\sum_{m=1}^{n} S(y_m)-S(\lambda)=C_{\text{tot}}(\lambda)\,. \qquad (2.32)$$

Thus one obtains the following theorem.

Theorem. The set of all variables in consideration is divided into subsets, and each subset is again subdivided into sub-subsets, et cetera, until finally the entire set is branched into individual variables. Then, the sum of all correlations, each of which is defined with respect to a branching point, is independent of the way in which this branching procedure is made and is equal to the total correlation.

This theorem can also be considered as a prescription for expanding the total correlation $C_{\text{tot}}(\lambda)$ in terms of the partial correlations of the type: $C(\mu;\nu_1,\nu_2,\ldots)$. We can write (2.32) in the form:

$$C_{\text{tot}}=\sum_{\text{all}} C_{\text{partial}}(\mu;\nu_1,\nu_2,\ldots)\,, \qquad (2.33)$$

where C_{partial} should be taken at every branching point in the "taxonomical tree" whose stem is λ and the peripheral branches are individual variables, y.

Figures 1 and 2 illustrate the theorem in a special case, in which λ consists of seven variables, $y_1, y_2, \ldots, y_7$. First λ is divided into three subsets: $\mu_1=(y_1,y_2,y_3,y_4)$, $\mu_2=(y_5,y_7)$ and $\mu_3=y_6$. Then μ_1 is divided into two subsets: $\nu_1=(y_1,y_2,y_3)$, $\nu_2=y_4$. Finally, ν_1 and μ_2 are subdivided into individual y's. Thus there are four branching points, at which the correlations are:

$$C_1=C(\lambda;\mu_1,\mu_2,\mu_3)=S(\mu_1)+S(\mu_2)+S(y_6)-S(\lambda)\,,$$

$$C_2=C(\mu_1;\nu_1,\nu_2)=S(\nu_1)+S(y_4)-S(\mu_1)\,,$$

$$C_3=C(\nu_1;y_1,y_2,y_3)=S(y_1)+S(y_2)+S(y_3)-S(\nu_1)\,,$$

$$C_4=C(\mu_2;y_5,y_7)=S(y_5)+S(y_7)-S(\mu_2)\,.$$

Now, the sum total becomes

$$C_1+C_2+C_3+C_4=S(y_1)+S(y_2)+S(y_3)+S(y_4)+S(y_5)$$

$$+S(y_6)+S(y_7)-S(\lambda)\,.$$

The left-hand side is the sum of correlation at all branching points in Fig. 1, while the right-hand side is the correlation at the single branching point in Fig. 2, which is, of course, $C_{\text{tot}}(\lambda)$.

One possible branching scheme is to split off one variable at one time. This means, conversely, that y_1 and y_2 are first grouped together to form a subset (y_1y_2) and then y_3 is added to form a higher subset $(y_1y_2y_3)$, et cetera. See Fig. 3. Then the theorem can be written as

$$C[(y_1y_2);y_1,y_2]+C[(y_1y_2y_3);(y_1y_2),y_3]+\ldots$$

$$\ldots+C[(y_1y_2\ldots y_n);(y_1,y_2\ldots y_{n-1}),y_n]$$

$$=C[(y_1y_2\ldots y_n);y_1,y_2,y_3,\ldots,y_n]=C_{\text{tot}}(\lambda)\,. \qquad (2.34)$$

A typical term $C[(y_1y_2\ldots y_r);(y_1y_2\ldots y_{r-1}),y_r]$ in this expansion can be interpreted as follows. Suppose a subset $(y_1,y_2,\ldots,y_{r-1})$ and the variable y_r are observed separately. Then the information carried by them are respectively $S(y_1,y_2,\ldots,y_{r-1})$ and $S(y_r)$. But the information carried simultaneously by all r variables $(y_1,y_2,\ldots,y_r)$ is $S(y_1,y_2,\ldots,y_r)$. The decrease in information is given by $C[(y_1y_2\ldots y_r);(y_1y_2\ldots y_{r-1}),y_r]$. Alternatively, in terms of "ignorance," the ignorance about the value of y_r before *any* observation is $S(y_r)$. Now, if one has observed the values of $y_1, y_2\ldots$, and y_{r-1}, then the average ignorance about the value of y_r is reduced to $S(y_1,y_2,\ldots,y_r)-S(y_1,y_2,\ldots,y_{r-1})$. See (2.18). This decrease, $S(y_r)-[S(y_1,y_2,\ldots,y_r)-S(y_1,y_2,\ldots,y_{r-1})]$, is the information about y_r provided by the observation of $y_1,\ldots,y_{r-1}$ and is equal to $C[(y_1,y_2,\ldots,y_r);(y_1,y_2,\ldots,y_{r-1}),y_r]$. Equation (2.34) shows that $C_{\text{tot}}(\lambda)$ can be expressed as the sum of these terms obtained by gradually increasing r from 2 to n.

Each term in the decomposition (2.32), or (2.33) of C_{tot} is guaranteed to be non-negative due to the inequality (2.11). We can decompose $S(\lambda)$ into terms, each of which is guaranteed to be non-negative by virtue of the inequality (2.20). Suppose we make a chain of subsets of variables $\mu_1,\mu_2,\ldots,\mu_k=\lambda$ such that

$$\mu_1<\mu_2<\ldots<\mu_{k-1}<\mu_k=\lambda\,, \qquad (2.35)$$

where the symbol $<$ means "$\subset$ but not $=$".

Then, due to (2.20) we have

$$T_i\equiv S(\mu_i)-S(\mu_{i-1})\geq 0\,. \qquad (2.36)$$

Therefore, we obtain a decomposition

$$S(\lambda)=S(\mu_1)+T_2+T_3+\ldots+T_k\,, \qquad (2.37)$$

of which each term is non-negative. T_i means the increase in information by observation of variables of μ_i in addition to the information obtained from variables of μ_{i-1}.

If we take

$$\mu_1=y_1,\ \mu_2=(y_1y_2),\ \mu_3=(y_1y_2y_3)\ldots$$

$$\ldots\mu_k=(y_1y_2y_3\ldots y_n)\,,\qquad (n=k)\,, \qquad (2.38)$$

then (2.37) becomes

$$S(\lambda) = S(y_1)$$
$$+[S(y_1, y_2) - S(y_1)]$$
$$+[S(y_1, y_2, y_3) - S(y_1, y_2)]$$
$$+ \text{----}$$
$$+[S(y_1, y_2, \ldots, y_n) - S(y_1 y_2, , y_{n-1})]. \quad (2.39)$$

Subtracting from both sides $\sum_{i=1}^{n} S(y_i)$, one obtains formula (2.34) with the negative sign.

3. Expansions in terms of average entropy functions

For any function $F(r)$ of an integer argument r, satisfying

$$F(r) = 0, \qquad \text{for } r \leq 0, \quad (3.1)$$

we can easily prove a simple mathematical theorem:

$$F(n) = \sum_{r=1}^{n} \binom{n-r+t-1}{t-1} [F(r)]_t \text{ for any } t \geq 1, \quad (3.2)$$

where $[F(r)]_t$ is the t^{th} difference defined by

$$[F(r)]_t = \frac{\Delta^t F(r)}{\Delta r^t} = \sum_{s=0}^{t} \binom{t}{s} (-1)^s F(r-s). \quad (3.3)$$

They are, for instance,

$$[F(r)]_0 = F(r),$$
$$[F(r)]_1 = F(r) - F(r-1),$$
$$[F(r)]_2 = F(r) - 2F(r-1) + F(r-2),$$
$$[F(r)]_3 = F(r) - 3F(r-1) + 3F(r-2) - F(r-3). \quad (3.4)$$

Another useful expansion of $F(n)$ is

$$F(n) = \sum_{r=1}^{n} \binom{n}{r} [F(r)]_r, \quad (3.5)$$

which is equally easy to prove.

Now, coming back to the problem of the preceding section, we define an "average r-signal information," $\overline{S}^{(r)}$ by

$$\overline{S}^{(r)} = \sum_{\mu \in \lambda} S^{(r)}(\mu) \Big/ \binom{n}{r}, \quad (3.6)$$

where μ is a subset with r variables, and the summation is taken over all $\binom{n}{r}$ different subsets of r variables taken out of the original set λ of n variables. It should be mentioned here that such an average can be defined in any arbitrary case, but its usefulness becomes important only in the cases where the individual $S^{(r)}$'s are not very different from the average.

Suppose we take a subset μ with r variables in λ, and subdivide this μ into two sub-subsets α and β, respectively, with s and t variables. Then, from the consideration of the last section, we have

$$S^{(r)}(\mu) \leq S^{(s)}(\alpha) + S^{(t)}(\beta), \quad \mu = \alpha \cup \beta, \quad r = s + t. \quad (3.7)$$

Figure 1 **Three-stage polychotomy of states.**

$$\lambda = \mu_1 \cup \mu_2 \cup \mu_3 = (\nu_1 \cup \nu_2) \cup \mu_2 \cup \mu_3$$
$$= ((y_1 \cup y_2 \cup y_3) \cup y_4) \cup (y_5 \cup y_7) \cup y_6$$
$$= y_1 \cup y_2 \cup y_3 \cup y_4 \cup y_5 \cup y_6 \cup y_7.$$

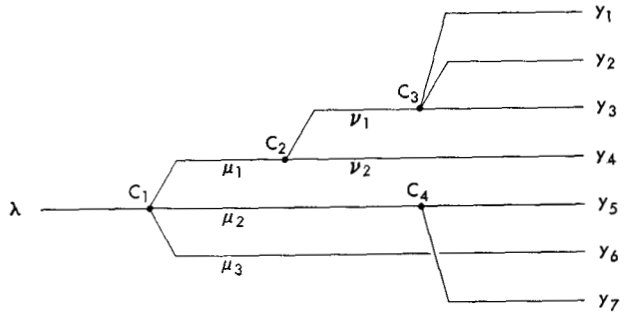


Figure 2 **One-stage polychotomy of states.**

$$\lambda = y_1 \cup y_2 \cup y_3 \cup y_4 \cup y_5 \cup y_6 \cup y_7.$$

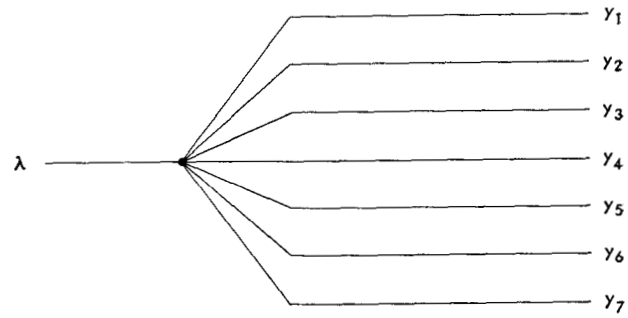
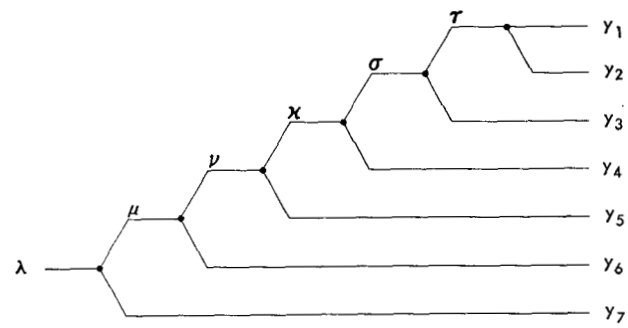


Figure 3 **Polychotomy by splitting of one state at each stage.**

$$\lambda = (\mu \cup y_7) = ((\nu \cup y_6) \cup y_7)$$
$$= (((\kappa \cup y_5) \cup y_6) \cup y_7)$$
$$= ((((\sigma \cup y_4) \cup y_5) \cup y_6) \cup y_7)$$
$$= (((((\tau \cup y_3) \cup y_4) \cup y_5) \cup y_6) \cup y_7)$$
$$= ((((((y_1 \cup y_2) \cup y_3) \cup y_4) \cup y_5) \cup y_6) \cup y_7).$$



Now, there are $\binom{r}{s}$ different ways to take (α, β) in μ. Thus

$$S^{(r)}(\mu) \leq \left[\sum_{\alpha \epsilon \mu} S^{(s)}(\alpha) \Big/ \binom{r}{s} + \sum_{\beta \epsilon \mu} S^{(t)}(\beta) \Big/ \binom{r}{t} \right]. \tag{3.8}$$

It should be noted that equality in (3.8) requires equality in all $\binom{r}{s}$ different relations of the type (3.7). Therefore, it is an extremely stringent requirement. Now, let us take the summation over all $\binom{n}{r}$ different μ's that can be taken in λ. Then,

$$\sum_{\mu \epsilon \lambda} S^{(r)}(\mu) \leq \sum_{\mu \epsilon \lambda} \left\{ \left[\sum_{\alpha \epsilon \mu} S^{(s)}(\alpha) \Big/ \binom{r}{s} \right] + \left[\sum_{\beta \epsilon \mu} S^{(t)}(\beta) \Big/ \binom{r}{t} \right] \right\}. \tag{3.9}$$

Now, let us fix our attention on a particular α in (3.9) and ask how many times this α will appear in the summation over $\binom{n}{r}$ different μ's. This is the number of μ's that includes this particular α. Hence, the answer is obviously $\binom{n-s}{r-s}$. The right-hand side of (3.9) will then become

$$\left[\binom{n-s}{r-s} \Big/ \binom{r}{s} \right] \sum_{\alpha \epsilon \lambda} S^{(s)}(\alpha) + \left[\binom{n-t}{r-t} \Big/ \binom{r}{t} \right] \sum_{\beta \epsilon \lambda} S^{(t)}(\beta).$$

In virtue of the fact

$$\binom{n}{s} \Big/ \binom{r}{s} = \frac{n(n-1)\ldots(n-s+1)}{r(r-1)\ldots(r-s+1)} = \binom{n}{r} \Big/ \binom{n-s}{r-s},$$

Eq. (3.9) becomes

$$\sum_{\mu \epsilon \lambda} S^{(r)}(\mu) \leq \left\{ \left[\binom{n}{r} \Big/ \binom{n}{s} \right] \sum_{\alpha \epsilon \lambda} S^{(s)}(\alpha) + \left[\binom{n}{r} \Big/ \binom{n}{t} \right] \sum_{\beta \epsilon \lambda} S^{(t)}(\beta) \right\},$$

or equivalently, in virtue of (3.6),

$$\overline{S}^{(r)} \leq \overline{S}^{(s)} + \overline{S}^{(t)}. \tag{3.10}$$

Thus, we can define an "average correlation":

$$\overline{C}(r; s, t) = \overline{S}^{(s)} + \overline{S}^{(t)} - \overline{S}^{(r)} \geq 0, \qquad r = s + t. \tag{3.11}$$

It should be noted, however, that in marked contrast to the case of an "individual" correlation, vanishing of an average correlation is a very strong condition.

The total correlation existing in a subset of r variables $y_1, y_2, \ldots, y_r)$ is, according to (2.29),

$$C^{(r)}_{\text{tot}}(y_1 y_2 \ldots y_r) = \sum_{i=1}^{r} S(y_i) - S(y_1 y_2 \ldots y_r). \tag{3.12}$$

The corresponding average total correlation of r-variable subsets is

$$\overline{C}^{(r)} = r\overline{S}^{(1)} - \overline{S}^{(r)}, \tag{3.13}$$

with

$$\overline{S}^{(1)} = \sum_{i=1}^{n} S(y_i)/n. \tag{3.14}$$

We now notice that we always have

$$\overline{S}^{(0)} = 0, \qquad \overline{C}^{(0)} = 0, \tag{3.15}$$

for $S^{(0)}$ is zero anyway. Since $\overline{S}^{(r)}$ and $\overline{C}^{(r)}$ are not defined for negative values of r, we can arbitrarily decide that they are zero. Thus,

$$\overline{S}^{(r)} = 0, \qquad \overline{C}^{(r)} = 0 \text{ for } r \leq 0, \tag{3.16}$$

in agreement with (3.1). Actually, we have, in addition to this, $\overline{C}^{(1)} = 0$. It must be kept in mind that the definition (3.13) is valid only for non-negative values of r. For $r = -1$, for instance, it would give $\overline{C}^{(-1)} = -\overline{S}^{(1)}$, which is not what is intended, since $C^{(-1)}$ is zero by definition due to (3.16).

Now, according to (3.6) and (3.13), we see that for $r = n$, the quantities $\overline{S}^{(r)}$ and $\overline{C}^{(r)}$ are not an "average," but the total information and total correlation in λ.

$$\overline{S}^{(n)} = S(\lambda) = n\overline{S}^{(1)} - C_{\text{tot}}(\lambda), \tag{3.17}$$

$$\overline{C}^{(n)} = C_{\text{tot}}(\lambda) = n\overline{S}^{(1)} - S(\lambda). \tag{3.18}$$

Therefore, substituting $\overline{S}^{(r)}$ or $\overline{C}^{(r)}$ for $F(r)$ in (3.2) and (3.5), we obtain all kinds of expansions of $S(\lambda)$ and $C_{\text{tot}}(\lambda)$ in terms of $\overline{S}^{(r)}$ or $\overline{C}^{(r)}$. We shall mention just a few of them as illustrations. For instance, for $n = 4$, $t = 4$, Eq. (3.2) gives, with $F(r) = \overline{S}^{(r)}$,

$$S(\lambda) = 20\overline{S}^{(1)} + 10(\overline{S}^{(2)} - 4\overline{S}^{(1)}) + 4(\overline{S}^{(3)} - 4\overline{S}^{(2)} + 6\overline{S}^{(1)} + (\overline{S}^{(4)} - 4\overline{S}^{(3)} + 6\overline{S}^{(2)} - 4\overline{S}^{(1)}). \tag{3.19}$$

An interesting expansion is obtained from (3.2) by putting $t = 1$ and identifying $F(r)$ with $\overline{C}^{(r)}$.

$$C_{\text{tot}}(\lambda) = \sum_{r=1}^{n} [\overline{C}^{(r)}]_1 = \sum_{r=2}^{n} (\overline{C}^{(r)} - \overline{C}^{(r-1)}) = \sum_{r=2}^{n} [\overline{S}^{(1)} - \overline{S}^{(r)} + \overline{S}^{(r-1)}]. \tag{3.20}$$

One can immediately recognize that (3.20) corresponds to the expansion (2.34) which we discussed in some detail in the last section. The only difference is that Eq. (3.20) is expressed in terms of "average" information, while Eq. (2.34) is expressed in terms of information of well-

defined variables, along a chain like one in Fig. 3.

Next, by putting $\bar{S}^{(r)}$ as $F(r)$ in (3.5), one obtains

$$S(\lambda)=n\bar{S}^{(1)}+\binom{n}{2}(\bar{S}^{(2)}-2\bar{S}^{(1)})+\binom{n}{3}(\bar{S}^{(3)}-3\bar{S}^{(2)}+3\bar{S}^{(1)})$$
$$+\binom{n}{4}(\bar{S}^{(4)}-4\bar{S}^{(3)}+6\bar{S}^{(2)}-4\bar{S}^{(1)})+\ldots. \quad (3.21)$$

By transferring the first term on the right side to the left side, one obtains an expression for $C_{\text{tot}}(\lambda)$. Each of the remaining terms on the right side has rather a nice symmetrical form, for instance,

$$\binom{n}{2}(\bar{S}^{(2)}-2\bar{S}^{(1)})=\Sigma[S(y_1y_2)-S(y_1)-S(y_2)] \quad (3.22)$$

$$\binom{n}{3}(\bar{S}^{3}-3\bar{S}^{2}+3\bar{S}^{1})=\Sigma[S(y_1y_2y_3)-S(y_1y_2)-S(y_2y_3)$$
$$-S(y_3y_1)+S(y_1)+S(y_2)+S(y_3)], \text{ et cetera.} \quad (3.23)$$

The summation in (3.22) is to be extended over all pairs of y's, and the summation in (3.23) is to be extended over all groups of three variables. The term in bracket in (3.22) is obviously $-C[(y_1y_2);y_1,y_2]\leq 0$ and represents the decrease of information due to correlation between two elements y_1 and y_2. It might be tempting to consider the quantity in the brackets in (3.23) as representing the decrease of information due to the correlation peculiar to a three-element system, $(y_1y_2y_3)$. This interpretation, however, is hardly justifiable since the terms in expansion (3.21) can be both positive and negative, although these quantities have a certain formal beauty, e.g., it can be written for three variables,

$$-\sum_{x_1,x_2,x_3}^{3} p(x_1,x_2,x_3)$$
$$\log\frac{p(x_1,x_2,x_3)p(x_1)p(x_2)p(x_3)}{p(x_1,x_2)p(x_2,x_3)p(x_3,x_4)}. \quad (3.24)$$

The reason why this expression does not have a profound meaning is that the quantity

$$\frac{p(x_1,x_2)p(x_2,x_3)p(x_3,x_1)}{p(x_1)p(x_2)p(x_3)}, \quad (3.25)$$

which occupies a position comparable to $p(x_1,x_2,x_3)$ in (3.24), cannot be considered as a probability distribution for (x_1,x_2,x_3) since it does not satisfy the normalization condition. The expansion of this type has been considered by Professor R. M. Fano (in an oral presentation at IBM Research Laboratory).

Consider for example, the case of four binary variables in which $p(x_1,x_2,x_3,x_4)$ is determined by $p(0,1,0,1)=p(1,0,1,0)=1/2$. Then we obtain $S^{(1)}=1$, $\bar{S}^{2}-2\bar{S}^{1}=-1$, $\bar{S}^{3}-3\bar{S}^{(2)}+3\bar{S}^{(1)}=+1$, $S^{4}-4\bar{S}^{(3)}+6\bar{S}^{(2)}-4\bar{S}^{(1)}=-1$, and the expansion in question (3.21) consists of four terms and is given by $S^{(4)}=-1+4-6+4=+1$. This alternation of signs does not allow of any useful interpretation. The expansions (2.32), (2.33), (2.34), (3.20) have the definite advantage that each term in the summation has a definite sign, contributing to the total correlation. The expansions such as (3.19) and (3.21) do not have this advantage.

Before passing to the next section, it may be worthwhile to give further consideration to the "average" correlations to clarify their meaning. To make the argument concrete, let us take the case $n=4$ and discuss the expansion (3.20) which may be written

$$C_{\text{tot}}^{(4)}=V^{(4)}+V^{(3)}+V^{(2)}, \quad (3.26)$$

with

$$V^{(r)}=\bar{S}^{(1)}+\bar{S}^{(r-1)}-\bar{S}^{(r)}\geq 0. \quad (3.27)$$

$V^{(r)}$ is a special case of (3.11) with $t=1$. To understand the meaning of $V^{(r)}$, it is helpful to know when it vanishes. In terms of "individual" correlations of the type (2.19), we have four relations of the form

$$C[(y_1y_2y_3y_4);(y_1y_2y_3),y_4]$$
$$=S^{(3)}(y_1y_2y_3)+S^{(1)}(y_4)-S^{(4)}(y_1y_2y_3y_4)\geq 0, \quad (3.28)$$

in which one out of four variables $(y_1y_2y_3y_4)$ is singled out. $V^{(4)}$ of (3.27) is the sum of these four relations, divided by four. Therefore, equality in (3.27) entails equality in each of the four relations of the type (3.28). This means that the condition

$$V^{(4)}=0 \quad (3.29)$$

implies, in virtue of (2.12),

$$p(x_1,x_2,x_3,x_4)=p(x_1x_2x_3)p(x_4)=p(x_2x_3x_4)p(x_1)$$
$$=p(x_3x_4x_1)p(x_2)=p(x_4x_1x_2)p(x_3), \quad (3.30)$$

which in turn means four equations of the type

$$p(x_1|x_2,x_3,x_4)=p(x_1). \quad (3.31)$$

Equation (3.31) seems to imply that all four variables are independent of one another. That this is indeed the case will presently be seen. By summing over x_4 in (3.30), one obtains

$$p(x_1x_2x_3)=p(x_2x_3)p(x_1)=p(x_3,x_1)p(x_2)$$
$$=p(x_1x_2)p(x_3). \quad (3.32)$$

By summing over x_1, x_2 and x_3, one obtains similar expressions for $p(x_2x_2x_3)$, $p(x_1,x_3,x_4)$ and $p(x_1,x_2,x_4)$. These equations give

$$V^{(3)}=0. \quad (3.33)$$

By the same token, one obtains from (3.33)

$$V^{(2)}=0, \quad (3.34)$$

or equivalently, six relations of the type

$$p(x_1,x_2)=p(x_1)p(x_2). \quad (3.35)$$

Substituting (3.35) in (3.32) and again substituting the resulting relations in (3.30), one obtains finally the relation of complete independence,

$$p(x_1x_2x_3x_4)=p(x_1)p(x_2)p(x_3)p(x_4)\,, \tag{3.36}$$

which means $C^{(4)}_{\text{tot}}=0$. This last relation can also be concluded by putting (3.29), (3.33), (3.34) in (3.26).

It is very important to note the relations of the type (3.35) *alone* do not imply (3.32) or (3.36). This means that even if $V^{(2)}=0$, it is quite possible that $V^{(3)}\neq 0$ and $V^{(4)}\neq 0$. What has been proven above is that if $V^{(4)}=0$, then $V^{(3)}=V^{(2)}=0$. This means, if $V^{(2)}\neq 0$, then $V^{(3)}\neq 0$ and $V^{(4)}\neq 0$. In general, nonvanishing, lower-range *average* correlations imply nonvanishing higher-range *average* correlation. This situation offers a contrast to the case of W's of the next section, where it is possible that the W's of higher range can vanish while the W's of lower range do not.

In any event, the concept of average correlations becomes useful only when all the variables $y_1, y_2 \ldots, y_n$ are more or less on the same footing as in the case of simultaneous signals. In particular, it is a very natural concept when the variables are absolutely equivalent in a statistical sense, i.e., when the probability distribution remains invariant for any permutation of values of the variables,

$$p(y_1=x_1, \ldots, y_i=x_i, \ldots, y_j=x_j, \ldots, y_n=x_n)$$
$$=p(y_1=x_1, \ldots, y_i=x_j, \ldots, y_j=x_i, \ldots, y_n=x_n)$$
for any pair (i, j). (3.37)

In this case, all the $\binom{n}{r}$ r-signal information functions are equal to one another and to the average. Thus, in such a case, $\bar{S}^{(r)}$ is not an average, but the actual r-signal information. This kind of thing can happen when we apply the present analysis to a physical system, such as a gas. The meaning of $V^{(r)}$ is obvious. If one observes any $(r-1)$ signals and any one signal separately, one obtains respectively information $S^{(r-1)}$ and $S^{(1)}$. But if one observes all these signals together, one obtains only $S^{(r)}$. The loss in information, i.e., correlation, is $V^{(r)}$. Expansion (3.20) or (3.26) can be interpreted as computation of the total correlation by adding these $V^{(r)}$, starting from one signal and increasing gradually the number of signals observed.

4. Correlation in stochastic time-sequence

We are given an infinite, one-dimensional (temporal) series of stochastic variables

$$\ldots, y_{-3}, y_{-2}, y_{-1}, y_0, y_1, y_2, y_3, \ldots, \tag{4.1}$$

each having the same domain of g values, such that any arbitrary segment of n consecutive variables has a unique and definite probability of having a given ordered set of values, say $(x_1, x_2, \ldots, x_n)$. This implies that

$$p(y_1=x_1, \ldots, y_n=x_n)=p(y_{1+k}=x_1, \ldots, y_{n+k}=x_n), \tag{4.2}$$

where k is an arbitrary integer, positive or negative. For this reason, we shall denote the probability given (4.2) simply by $p(x_1, \ldots, x_n)$, or still more simply $p^{(n)}$. We can theoretically divide these n variables in any way we wish, and consider the subsets thus produced. All the formulas of Sections 2 and 3 will also apply to this case. We are not particularly interested, however, in a subset consisting of variables scattered here and there on the one-dimensional line. We are interested in a subset consisting of consecutive variables. The formulas of Section 3 in terms of average entropies for two variables, three variables, et cetera, are not interesting here, since in the present case there is a clear definition of distance between two variables. The relation between y_1 and y_2 is thus entirely different, say, from the relation between y_1 and y_{100}.

Once the probability $p^{(n)}$ for n consecutive variables is given, then all the probabilities $p^{(r)}$ for $r\leq n$ consecutive variables can be obtained by the use of the recurrence formula:

$$p^{(n-1)}(x_1, x_2, \ldots, x_{n-1})=\sum_{x_n} p^{(n)}(x_1, x_2, \ldots, x_{n-1}, x_n)$$
$$=\sum_{x_0} p^{(n)}(x_0, x_1, \ldots, x_{n-1})\,. \tag{4.3}$$

On account of (4.2), the information function of a segment of length r

$$S^{(r)}=-\sum_{x}{}^{r} p^{(r)}(x_1, \ldots, x_r)\log p^{(r)}(x_1, \ldots, x_r) \tag{4.4}$$

depends only on r. Similarly, the total correlation existing in a segment of length

$$C^{(r)}=rS^{(1)}-S^{(r)} \tag{4.5}$$

depends only on r.

As a consequence, we can use formulas (3.2) and (3.5) to obtain different expansions of $C^{(n)}$ or $S^{(n)}$. Particularly useful are the following three. First, by identifying $F(r)$ with $S^{(r)}$ in (3.2) with $t=1$, one obtains

$$nS^{(1)}-C^{(n)}_{\text{tot}}=S^{(n)}=\sum_{r=1}^{n}[S^{(r)}]_1=S^{(1)}+(S^{(2)}-S^{(1)})$$
$$+(S^{(3)}-S^{(2)})+\ldots+(S^{(n)}-S^{(n-1)})\,. \tag{4.6}$$

Identifying $F(r)$ with $C^{(r)}$ in (3.2) with $t=1$, one obtains

$$nS^{(1)}-S^{(n)}=C^{(n)}_{\text{tot}}=\sum_{r=1}^{n}[C^{(r)}]_1=(2S^{(1)}-S^{(2)})$$
$$+(S^{(1)}-S^{(3)}+S^{(2)})+\ldots+(S^{(1)}-S^{(r)}+S^{(r-1)})$$
$$+\ldots+(S^{(1)}-S^{(n)}+S^{(n-1)})\,, \tag{4.7}$$

which looks similar to (3.15) but it is very important the S's here are not the average but the information function of a segment of given length. Finally, identifying $F(r)$ with $C^{(r)}$ with $t=2$, one obtains

$$nS^{(1)}-S^{(n)}=C^{(n)}_{\text{tot}}=\sum_{r=1}^{n}(n-r+1)[C^{(r)}]_2$$
$$=(n-1)C^{(2)}+(n-2)(C^{(3)}-2C^{(2)})$$
$$+(n-3)(C^{(4)}-2C^{(3)}+C^{(2)})+\ldots$$

$$+(C^{(n)}-2C^{(n-1)}+C^{(n-3)})$$
$$=(n-1)(2S^{(1)}-S^{(2)})+(n-2)(-S^{(1)}$$
$$+2S^{(2)}-S^{(3)})+\ldots$$
$$+(-S^{(n-2)}+2S^{(n-1)}-S^{(n)}). \qquad (4.8)$$

Introducing the "correlation indices" $W^{(r)}$ by[13]

$$W^{(r)}=-S^{(r)}+2S^{(r-1)}-S^{(r-2)}, \qquad (4.9)$$

one can write the total correlation and the information per position as

$$C_{\text{tot}}^{(n)}=\sum_{r=2}^{n}(n-r+1)W^{(r)} \qquad (4.10)$$

$$I^{(n)}\equiv S^{(n)}/n=S^{(1)}-\sum_{r=2}^{n}[(n-r+1)/n]W^{(r)}$$

$$=\sum_{r=1}^{n}[(n-r+1)/n]W^{(r)}. \qquad (4.11)$$

In the last expression, the relation $W^{(1)}=-S^{(1)}$ is used. Note: $S^{(0)}=0$.

Now let us clarify the meaning of three expansions (4.6), (4.7) and (4.8) which we have just obtained. In the first place, the representative term $S^{(r)}-S^{(r-1)}$ in (4.6) is non-negative because of the relation (2.20). Next, λ and ν are here a sequence of length r and a sequence of length $(r-1)$. The letter μ stands for one symbol. Therefore, the relation $S^{(r)}-S^{(r-1)}=S(\lambda)-S(\nu)=0$ can be interpreted as meaning that the knowledge of the first $(r-1)$ positions in the sequence completely determines the last position. In other words, $S^{(r)}-S^{(r-1)}$ measures the further information carried by the last position over and above the information already given by the first $(r-1)$ positions. Equation (4.6) can also be obtained by calculating $S^{(n)}$ with the help of the formula

$$p^{(n)}(x_1,x_2,\ldots,x_n)=p^{(1)}(x_1)p(x_2|x_1)p(x_3|x_1x_2)\times$$
$$p(x_4|x_1x_2x_3)\ldots p(x_n|x_1x_2\ldots x_{n-1}). \qquad (4.12)$$

But this expansion is not particularly interesting since $S^{(r)}-S^{(r-1)}$ seldom becomes very small. When there is no correlation whatsoever, $S^{(r)}-S^{(r-1)}$ will become equal to $S^{(1)}$. This expansion has often been used by Shannon.[6]

Now, turning to (4.7), the representative term $-S^{(r)}+S^{(r-1)}+S^{(1)}$ is non-negative in virtue of (2.19). This expansion is exactly the same as (2.34) if $(y_1y_2\ldots y_n)$ in (2.34) are taken in the chronological order in a time-sequence. The meaning has already been studied.

Finally, the representative term $W^{(r)}$ in (4.10) is also non-negative for $r\geq 2$. This can easily be seen by considering a probability distribution

$$q^{(r)}(x_1,x_2,\ldots,x_r)=p(x_1,x_2,\ldots,x_{r-1})$$
$$\times\frac{p(x_2,x_3,\ldots,x_r)}{p(x_2,x_3,\ldots,x_{r-1})}. \qquad (4.13)$$

Then, the quantity

$$+(\sum_x)^r p^{(r)}(x_1,x_2,\ldots,x_r)\log p^{(r)}(x_1,x_2,\ldots,x_r)$$
$$-(\sum_x)^r p^{(r)}(x_1,x_2,\ldots,x_r)\log q^{(r)}(x_1,x_2,\ldots,x_r)$$
$$=-S^{(r)}+2S^{(r-1)}-S^{(r-2)}=W^{(r)} \qquad (4.14)$$

is, in virtue of the Gibbs theorem, non-negative and becomes zero if and only if $p^{(r)}$ and $q^{(r)}$ are equal to each other for all possible values of $(x_1,x_2,\ldots,x_r)$. If this is the case then, from (4.13), we have

$$\frac{p(x_1,x_2,\ldots,x_r)}{p(x_1,x_2,\ldots,x_{r-1})}=\frac{p(x_2,x_3,\ldots,x_r)}{p(x_2,x_3,\ldots,x_{r-1})} \qquad (4.15)$$

or equivalently

$$p(x_r|x_1,x_2,\ldots,x_{r-1})=p(x_r|x_2,x_3,\ldots,x_{r-1}). \qquad (4.16)$$

This means that if we predict x_r on the basis of the knowledge of $(x_1,x_2,\ldots,x_{r-1})$ or the knowledge of $(x_2,x_3,\ldots,x_{r-1})$ it does not make any difference. That means that the knowledge of x_1 which is r positions prior to x_r, over and above the knowledge of the in-between positions does not affect the conditional probability about the state of x_r. In this sense, (4.15) means an absence of correlation of range r over and above the correlation of range $(r-1)$. Therefore, $W^{(r)}$ can be considered as a measure of the strength of correlation of range r. Eq. (4.10) shows that the total correlation can be written as a sum of W's with suitable non-negative coefficients.

If there were no correlation at all, one would have $I^{(n)}=S^{(1)}$ in (4.11). Therefore, each term (which is non-negative) for $r\geq 2$ under the summation in (4.11) represents the loss of information per position due to the correlation or redundancy of range r. We can also write (4.11) as

$$I^{(n)}=S^{(1)}-\frac{1}{n}[W^{(n)}+2W^{(n-1)}+\ldots+(n-1)W^{(2)}]. \qquad (4.17)$$

The information per position in an infinite sequence is given by

$$I^{(\infty)}=\lim_{n\to\infty}I^{(n)}=\lim_{n\to\infty}S^{(n)}/n, \qquad (4.18)$$

where $I^{(n)}$ should be taken from (4.17). If there is an integer $m(\geq 1)$ such that

$$W^{(k)}=0 \text{ for } k>m, \qquad (4.19)$$

we can write

$$I^{(\infty)}=S^{(1)}-\sum_{r=2}^{\infty}W^{(r)} \qquad (4.20)$$

for this series will break off at a certain place.

If (4.19) is the case, one can also easily show

$$S^{(n)}=(n-m+1)S^{(m)}-(n-m)S^{(m-1)},$$
$$I^{(\infty)}=S^{(m)}-S^{(m-1)}. \qquad (4.21)$$

If the chain is Markovian, then (4.19) holds for $m=2$.

(Usually, $W^{(2)} \neq 0$.) In this case one has from (4.20) and (4.21),

$$I^{(\infty)} = S^{(2)} - S^{(1)}, \tag{4.22}$$

$$S^{(n)} = (n-1)S^{(2)} - (n-2)S^{(1)}. \tag{4.23}$$

Another way to look at the situation is to note that

$$\frac{\Delta S^{(r)}}{\Delta r} = S^{(r)} - S^{(r-1)} \geq 0 \tag{4.24}$$

because of (2.20), and that

$$\frac{\Delta^2 S^{(r)}}{\Delta r^2} = S^{(r)} - 2S^{(r-1)} + 2S^{(r-2)} \leq 0 \tag{4.25}$$

because of (4.14). This shows that $S^{(r)}$ as a function of r is a monotonously increasing curve with a gradually decreasing (convex) slope. There is, therefore, a limiting slope. Particularly, if (4.19) is true, then the slope remains constant for $r > m$. The slope, meaning the information increase per position, will become $I^{(\infty)}$ after it has reached its final value. Thus, $I^{(\infty)} = S^{(m)} - S^{(m-1)}$ is easily understandable.

It is instructive to interpret the formulae thus obtained once again from the point of view of the degree of our ignorance regarding the state of a position. We take any one position in the infinite sequence, say the 0^{th} position, and propose to guess the state of this position. Then, the ignorance (or degree of uncertainty) is given by

$$\text{Ign}^{(1)} = S^{(1)} = -\sum_{x_0=1}^{g} p(x_0) \log p(x_0). \tag{4.26}$$

Next, suppose that we know already that the preceding position, i.e., the $(-1)^{\text{st}}$ position, was in a certain state, say, x_{-1}. Then, our ignorance regarding the state of the 0^{th} position becomes

$$-\sum_{x_0=1}^{g} p(x_0|x_{-1}) \log p(x_0|x_{-1})$$

$$= -\sum_{x_0=1}^{g} \frac{p(x_{-1}, x_0)}{p(x_{-1})} \log \frac{p(x_{-1}, x_0)}{p(x_{-1})}. \tag{4.27}$$

However, the $(-1)^{\text{st}}$ position turns out to be in the state x_{-1} with probability $p(x_{-1})$. Therefore, the ignorance about the state of the 0^{th} position, on the basis of the knowledge of the state of the $(-1)^{\text{st}}$ position, becomes on the average

$$\text{Ign}^{(2)} = -\sum_{x_{-1}=1}^{g} p(x_{-1}) \sum_{x_0=1}^{g} \frac{p(x_{-1}, x_0)}{p(x_{-1})} \log \frac{p(x_{-1}, x_0)}{p(x_{-1})}$$

$$= S^{(2)} - S^{(1)}. \tag{4.28}$$

Similarly, the ignorance about the state of the 0^{th} position on the basis of the knowledge about the states of r preceding positions, i.e., $(-1)^{\text{st}}$, $(-2)^{\text{nd}}$, . . . , and $(-r)^{\text{th}}$ positions will be

$$\text{Ign}^{(r+1)} = S^{(r+1)} - S^{(r)}. \tag{4.29}$$

The decrease in ignorance, i.e., increase in knowledge, about the state of the 0^{th} position by knowing one more position in the past is given by

$$\text{Ign}^{(r)} - \text{Ign}^{(r+1)} = -S^{(r+1)} + 2S^{(r)} - S^{(r-1)} = W^{(r+1)}. \tag{4.30}$$

This derivation may serve to give further clarification of the meaning of the correlation index $W^{(r)}$. The condition $W^{(r+1)} = 0$ means that the additional knowledge about the $(-r)^{\text{th}}$ position does not change our prediction about the 0^{th} position. The more one learns about the past, the less our ignorance about the present becomes. The original ignorance is $S^{(1)}$, the second-stage ignorance is $S^{(1)} - W^{(2)}$, and the third-stage ignorance is $S^{(1)} - W^{(2)} - W^{(3)}$, et cetera. Thus the minimum ignorance is

$$(\text{Ign})_{\min} = S^{(1)} - W^{(2)} - W^{(3)} - \cdots\cdots. \tag{4.31}$$

This ignorance represents the uncertainty in guessing the state of the 0^{th} position with the best knowledge of the past. The moment the observation determines this state, this uncertainty disappears, i.e., the ignorance becomes zero. This decrease in ignorance is the so-called "information." Thus, it is not unexpected that of (4.20) and $(\text{Ign})_{\min}$ of (4.31) coincide. $W^{(r)}$ represents, on one hand, the decrease of information due to correlation of range r, and, on the other hand, the average increase of our knowledge about the state of a position by knowing the state r positions prior to it over and above the knowledge about the states of the in-between positions.

It is often stated that if the correlation is of a finite range, then one can take segments of sufficient length and treat them as if they were independent. We shall here give a formula by which one can evaluate the error committed by such a procedure. Suppose the range is m or less, i.e., $W^{(k)} = 0$ for $k > m$, and take two consecutive segments of lengths, n_1 and n_2, such that $n_1 > m$, $n_2 > m$. The information carried by the entire segment of length $n = n_1 + n_2$, by the segment of length n_1 and by the segment of length n_2 are, respectively,

$$S^{(n_1+n_2)} = (n_1+n_2)S^{(1)} - \sum_{r=2}^{m} (n_1+n_2-r+1) W^{(r)},$$

$$S^{(n_1)} = n_1 S^{(1)} - \sum_{r=2}^{m} (n_1-r+1) W^{(r)},$$

$$S^{(n_2)} = n_2 S^{(1)} - \sum_{r=2}^{m} (n_2-r+1) W^{(r)}. \tag{4.32}$$

Therefore the correlation omitted in the above procedure is

$$S^{(n_1)} + S^{(n_2)} - S^{(n_1+n_2)} = \sum_{r=2}^{m} (r-1) W^{(r)}. \tag{4.33}$$

The ratio of the quantity in (4.33) to $S^{(n_1+n_2)}$ in (4.32) gives the fractional error. This fraction becomes, of course, zero as $(n_1+n_2) \to \infty$, since the correlation (4.33) does not depend on n_1, n_2, or n, which is obvious because the correlation exists in the present case, in the vicinity of the border of the two segments.

5. Application 1: Redundancy in geometrical figures

In the problem of pattern recognition, it is often important to determine the degree of redundancy, i.e., the degree of possibility of guessing the entire figure by observing only part of it. The following extremely simple example may be sufficient to show that the correlation analysis, as developed in Section 2, can be a useful instrument in unraveling this type of problem.

Suppose one has a square which is divided into four equal square cells (see Fig. 4). Each cell can be black or white. Then there are $2^4=16$ different figures. In order to identify each one of them, let us introduce four variables y_1, y_2, y_3, y_4, corresponding to the four cells, in the order of left-upper, right-upper, left-lower and right-lower cells. Each variable can be 0 or 1, "0" meaning white and "1" meaning black, Thus, for instance, ($y_1=1$, $y_2=1$, $y_3=0$, $y_4=0$) or simply (1100) means a horizontal line in the upper row. If all the 16 figures (letters) are used, there is no redundancy, for partial knowledge of the figure does not help to any extent to identify the figure. But if the alphabet is limited to a fraction of 16 figures, redundancy appears.

To fix our idea, let us assume that the alphabet consists of four letters (1100), (1010), (0011) and (0101), i.e., upper horizontal line, left vertical line, lower horizontal line and right vertical line, and that they are used with equal probability:

$$p(1,1,0,0)=p(1,0,1,0)=p(0,0,1,1)$$
$$=p(0,1,0,1)=1/4\,. \tag{5.1}$$

Then obviously we have the total entropy

$$S(y_1,y_2,y_3,y_4)=2\,, \tag{5.2}$$

corresponding to the fact each letter can emit 2 bits of information. In terms of "ignorance," this means that before observation four cases are equally probable and we have not the slightest idea as to which of the four will appear. Similarly, the entropy functions of three variables are all equal to 2, since four different configurations of three cells can appear with equal probabilities:

$$S(y_2,y_3,y_4)=S(y_1,y_3,y_4)=S(y_1,y_2,y_4)$$
$$=S(y_1,y_2,y_3)=2\,. \tag{5.3}$$

But the entropy functions of two variables divide themselves into two categories. If the two variables are taken horizontally or vertically, their entropies are 2.

$$S(y_1,y_2)=S(y_3,y_4)=S(y_1,y_3)=S(y_2,y_4)=2\,. \tag{5.4}$$

Eq. (5.4) is true because there are four possible cases, (black-white), (white-black), (black-black) and (white-white), with equal probability. But if the two variables are taken diagonally their entropies are 1:

$$S(y_1,y_4)=S(y_2,y_3)=1\,. \tag{5.5}$$

This is true because, there are only two possible cases, (black-white) and (white-black). The entropy of one variable is obviously 1, since black and white appear with equal probability in each cell:

$$S(y_1)=S(y_2)=S(y_3)=S(y_4)=1\,. \tag{5.6}$$

Now the total correlation is

$$C_{\text{tot}}=S(y_1)+S(y_2)+S(y_3)+S(y_4)-S(y_1y_2y_3y_4)$$
$$=1+1+1+1-2=2\,. \tag{5.7}$$

This corresponds to the fact that with $p(0)=p(1)=1/2$, one could at best send four bits of information (using the 16 alphabets with equal probabilities). By limiting oneself to the four letters, however, one sends only two bits. The difference $4-2=2$ is the loss of information, or redundancy.

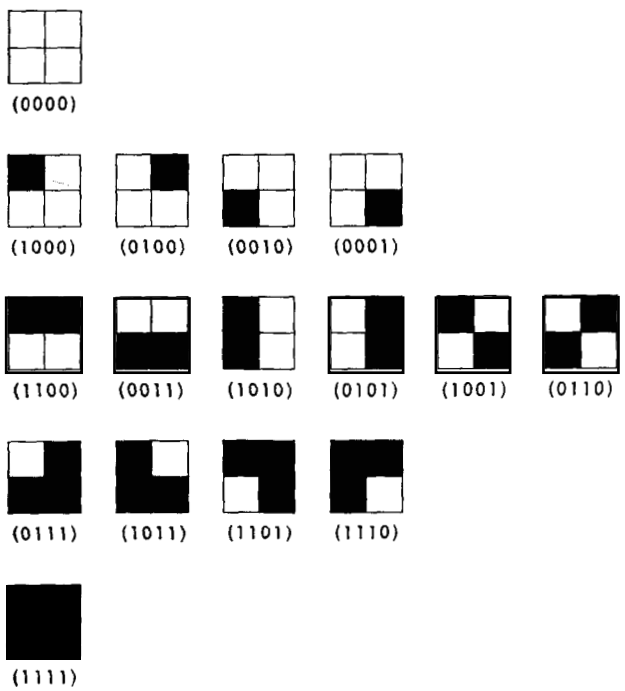
Now if we divide the set of four variables (y_1,y_2,y_3,y_4) into a group of three variables and one remaining variable, then the correlation corresponding to this "branching" is independent of the way the division is made and equal to

$$C[(y_1y_2y_3y_4);(y_1y_2y_3),y_4]$$
$$=2+1-2=1\,. \tag{5.8}$$

This means that a group of three variables convey two bits of information, while one variable conveys one bit of information, but all together they can convey still only

Figure 4 **Redundancy in geometrical figures.**

Simplified example of an application of correlation analysis to problems of pattern recognition.



two bits of information. This is because variable y_4 is completely determined by the other three variables. This case should be compared with formulae (2.21) and (2.22), identifying

$$\lambda=(y_1y_2y_3y_4)$$
$$\mu=(y_1y_2y_3) \tag{5.9}$$
$$\nu=y_4\,.$$

One sees that equality in (2.22) is holding in this case, meaning that ν is completely dependent on μ. However, equality in (2.21) is not holding here, because μ is not completely determined by ν.

More interesting is the case of division of four variables into two groups of two variables. There are two cases typified by

$$C[(y_1y_2y_3y_4);(y_1y_2),(y_3y_4)] =2+2-2=2 \tag{5.10}$$

and

$$C[(y_1y_2y_3y_4);(y_1y_4),(y_2y_3)] =1+1-2=0\,. \tag{5.11}$$

The first case (5.10) means, in the light of (2.21) and (2.22), that (y_1y_2) and (y_3y_4) are mutually completely dependent. If (y_1y_2) is (black-black), then (y_3y_4) is (white-white), and vice versa. If (y_1y_2) is (black-white), then (y_3y_4) is also (black-white), and vice versa, et cetera. In the second case (5.11), each group (y_1y_4) or (y_2y_3) means a diagonal and can be (black-white) or (white-black). And even if we know that (y_1y_4) is one of the two, say, (black-white), there is still probability 1/2 for (y_2, y_3) being (black-white) and probability 1/2 for (y_2, y_3) being (white-black). Therefore, this is the case of complete independence. Case (5.10) is the maximum correlation and (5.11) the minimum correlation.

Observation of one cell gives only one bit of information. Therefore it is sufficient to select two out of four possibilities but not sufficient to identify the figure. Observation of two cells diagonally placed gives also only one bit, therefore not sufficient to identify the entire figure. But, observation of two cells horizontally or vertically placed gives two bits of information and is sufficient to identify the figure. This is obvious from the illustration, but it is interesting to see the mathematical expression of the situation in the foregoing formulas.

One can write various decompositions, according to (2.32), (2.34) or (2.39), and it is instructive to understand the meaning of each term. However, since it is rather elementary, we shall not describe them here.

It is also interesting to see the difference between the present choice of four letters (1100), (1010), (0011), (0101) and another choice of four letters, say, (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), and (0, 0, 0, 1), each of which has one black and three white. If the frequency of each of these four is 1/4, the total information is 2, as before. The total correlation here is 1.24512, while the total correlation in the former case was 2. In the present case, the information carried by one cell is 0.81128, as compared with 1 in the former case. Hence, observation of one cell is less powerful here in guessing the entire figure than in the former example. Similarly, the information carried by two cells is 1.5, no matter which two are taken, i.e., it is insufficient to identify the figure. In the former example, any two cells horizontally or vertically laid carried two bits, sufficient to identify the figure.

In this section we discussed only the case where the probability of each letter is equal, but our instruments of Section 2 are devised so as to cope with more general cases. It will be useful to consider the coding problem of geometrical figures, taking advantage of the redundancy existing in the figures, and assuming a special type of deformations likely to occur in geometrical figures.

6. Checking of randomizing effect of shuffling

The correlation index $W^{(r)}$ is not a useful instrument when there is a correlation of a very large range. For instance, if a stochastic chain has a hidden periodicity of very long range, one will have to calculate $S^{(r)}$ and $W^{(r)}$ for very large r's to discover such a periodicity. Calculation of this kind is impracticable, since the number of terms to be added in this calculation increases exponentially with r. On the other hand, most of the stochastic chains studied in scientific and engineering problems have a small range of correlation. In such a case the correlation index becomes a powerful tool. As an illustration, we shall describe here an experiment made on the IBM 704 simulating shuffling of cards, in which the correlations are gradually destroyed as a result of shuffling. The cards bear the designation "1" or "0".

The original sequence of binary numbers is a stochastic chain, but correlation is built in, so that probability is very high that a segment arbitrarily taken from it has a pattern:

$$\ldots\ldots 0\,0\,0\,0\,1\,1\,1\,1\,0\,0\,0\,0\,1\,1\,1\,1\,0\,0\,0\,0\,1\,1\,1\,1\ldots\ldots \tag{6.1}$$

Let us for a moment assume that the sequence had a rigid rule (6.1). Suppose we give ourselves a task of guessing the value x_0 of the variable y_0. Without any preliminary knowledge of other digits, there is an equal probability for $x_0=0$ and for $x_0=1$. Now, suppose we know that y_{-1} was 1, i.e., $x_{-1}=1$. Then, this position y_{-1} may be, with equal probability, any one of the four possible positions in a run of four 1's. If y_{-1} is the first, second or third position in this run, then x_0 will be 1. If y_{-1} is the last position of the four, then x_0 will be 0. Hence, the probability of x_0's being 1 is now 3/4 and the probability of being 0 is 1/4. If we know x_{-1} and x_{-2}, then our prediction of x_0 will become more accurate. Finally, if we know x_{-1}, x_{-2}, x_{-3} and x_{-4}, then the prediction of x_0 is no longer probabilistic, but deterministic. And further knowledge of x_{-5} will no longer change our prediction about x_0. Therefore, $W^{(6)}=0$. And also, $W^{(r)}=0, \qquad r>6$.

Of course, a sequence which strictly obeys the rule (6.1) is not a *stationary* stochastic sequence in the sense of (4.2). The sequence we used, therefore, was produced by the following probabilistic rules determined by conditional probabilities of range 5:

$p(x_r|x_{r-4}, x_{r-3}, x_{r-2}, x_{r-1})$

given below. In the following, ε is a constant very small compared with unity.

$$\begin{aligned}
&p(1|0000)=1-\varepsilon \quad , \quad p(1|1000)=\varepsilon \\
&p(1|0001)=1-\varepsilon \quad , \quad p(1|1001)=1-\varepsilon \\
&p(1|0010)=\varepsilon \quad , \quad p(1|1010)=\varepsilon \\
&p(1|0011)=1-\varepsilon \quad , \quad p(1|1011)=1-\varepsilon \\
&p(1|0100)=\varepsilon \quad , \quad p(1|1100)=\varepsilon \\
&p(1|0101)=1-\varepsilon \quad , \quad p(1|1101)=1-\varepsilon \\
&p(1|0110)=\varepsilon \quad , \quad p(1|1110)=\varepsilon \\
&p(1|0111)=1-\varepsilon \quad , \quad p(1|1111)=\varepsilon
\end{aligned} \tag{6.2}$$

This table is so devised that no matter where a run starts, it has a high probability of continuing to length 4. The probability of a run's ending at a position does not depend on what happened before the run has started. More precisely, if a run has lasted for a length r less than 4, $(r=1, 2, 3)$, then the probability of its continuing one more place is $1-\varepsilon$. If a run has lasted four or more places, then the probability of its terminating by the appearance of the next place is $1-\varepsilon$.

The numbers are produced by conditional probabilities of range 5, which means that the probability of a position taking a certain value is determined by four preceding positions, and that the knowledge of one more position preceding these four does not change this probability. Therefore, according to the analysis of the preceding Section, $W^{(6)}$ should be zero.

The shuffling has been done in the following way. Suppose we have a sequence of length N. We divide this sequence in n segments in a certain random fashion which will presently be explained. Then we have the first, second, . . . , $(n-1)^{\text{th}}$, and n^{th} segments. We reverse the order and form a new sequence of length N, which is formed by putting the n segments in the order of the n^{th}, $(n-1)^{\text{th}}$, . . . , second and first segments taken from the original sequence. This entire process will be called *one shuffle*.

The method of division is as follows. We produce a sequence of binary numbers by the following Markovian conditional probabilities, $p(x_k|x_{k-1})$, which are

$$\begin{aligned}
&p(1|0)=\eta \, , \\
&p(1|1)=0 \, ,
\end{aligned} \tag{6.3}$$

where η is small compared with unity. This means that a run of 0's terminates with the rule of accidental death, while a run of 1's always has length one. Now every time "1" appears at one position, say at the position k, then the original sequence will cut between the position k and $k+1$. The average length of a segment is $(1/\eta)+1$, if N is very large. The number n of segments in a sequence of length N is a stochastic variable here.

The first task in our experiment on the IBM 704 consisted of producing $N=105{,}000$ digits of binary numbers obeying (6.2). At each situation, i.e., whenever the four preceding numbers are given, there are, according to (6.2), two kinds of events: one, (A), with probability ε, and the other, (B), with probability $1-\varepsilon$. We produced each time a random number of appropriate length between 0 and 1, and if the number was between 0 and ε, we took the event of class (A), and if the number was between ε and 1, we took the event of class (B). For the value $\varepsilon=0$, we obtain the regular sequence (6.1). If $\varepsilon=1/2$, then the whole process will become completely random. The case $\varepsilon=1$ does not give rise to a stationary Markov chain, but if the starting sequence is (0101), then it continues to produce a chain with 0 and 1 appearing alternately. When we speak of the case $\varepsilon=1$ in the following, we shall mean this special case.

If $\varepsilon=0$, then, without shuffling, we should have (with $N\to\infty$)

$$\begin{aligned}
&W^{(2)}=0.18872 \, , \\
&W^{(3)}=0.12256 \, , \\
&W^{(4)}=0.18872 \, , \\
&W^{(5)}=0.5 \, , \\
&W^{(6)}=0 \, ,
\end{aligned} \tag{6.4}$$

and the total correlation per digit (for $N\to\infty$)

$$\sum_{r=2}^{6} W^{(r)} = \sum_{r=2}^{\infty} W^{(r)} = \lim_{n\to\infty} C_{\text{tot}}^{(n)}/n = 1 \, . \tag{6.5}$$

For $\varepsilon=1/2$, we should have

$$W^{(2)}=W^{(3)}=W^{(4)}=W^{(5)}=W^{(6)}=0 \, , \quad \sum_{r=2}^{6} W^{(r)}=0 \, . \tag{6.6}$$

For $\varepsilon=1$, we should have

$$W^{(2)}=1 \, , \qquad W^{(3)}=W^{(4)}=W^{(5)}=W^{(6)}=0 \, , \quad \sum_{r=2}^{6} W^{(r)}=1 \, . \tag{6.7}$$

In actual experiments, we used $\varepsilon=1$, which corresponds to (6.7), and $\varepsilon=1/4$, $1/8$, $1/16$, which lie between (6.5) and (6.6).

The shuffling sequence (6.3) was produced by a method similar to the one used in producing the main chain of numbers. The constant η was chosen to be $1/16$ in every experiment. We denote hereafter the number of shuffles by ρ. We give here samples of a segment of 32 digits arbitrarily taken from the actual chain of 105,000 numbers, before shuffling ($\rho=0$), after three shuffles ($\rho=3$),

and after six shuffles ($\rho=6$), with $\varepsilon=/16$ and $\eta=1/16$.

$\rho=0$: 11110000111100001111000011111 0 (6.8)

$\rho=3$: 11110000111011000011110011100001 (6.9)

$\rho=6$: 00000110000111101000100111011100 (6.10)

With $\rho=0$, there is one irregularity in the 32 digits as indicated by an arrow in (6.8). With $\rho=3$, there are four irregularities, and with $\rho=6$, there are ten irregularities. By an irregularity is meant a deviation from the rule (6.2) with $\varepsilon=0$.

In our experiments, we first determined various $p^{(r)}$'s from the actual frequencies of the varieties of segments, and then calculated each $W^{(r)}$ ($r=2, 3, 4, 5, 6$) and the total correlation $\sum_{r=2}^{6} W^{(r)}$ which should become $\lim_{n\to\infty} C_{\text{tot}}^{(n)}/n$ in the limiting case $N\to\infty$. We tried $\varepsilon=1$, $1/4$, and $1/8$ for $\rho=0$, 1, 2, 3. For $\varepsilon=1/16$, we computed for $\rho=0$, 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10.

Figure 5 shows how the total correlation $\sum_{r=2}^{6} W^{(r)}$ decreases with the number ρ of shufflings for $\varepsilon=1$, $1/4$, $1/8$, $1/16$ in the range $\rho=0$ to $\rho=3$. Figure 6 shows how the

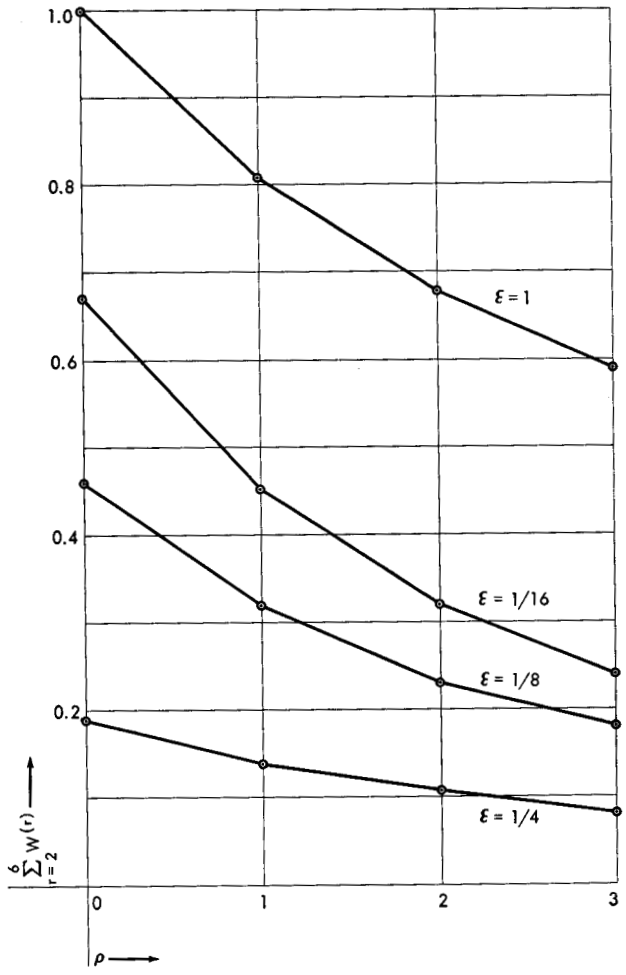


Figure 5 **Destruction of orderliness by three shuffles of cards marked 1 or 0.**

The total correlation per digit $\sum_{r=2}^{6} W^{(r)}$ *for* $\varepsilon=1$, $1/4$, $1/8$, $1/16$ *and for* $\rho=0$, 1, 2, 3. *This shows how the total correlation decreases with increasing number of shuffles.*

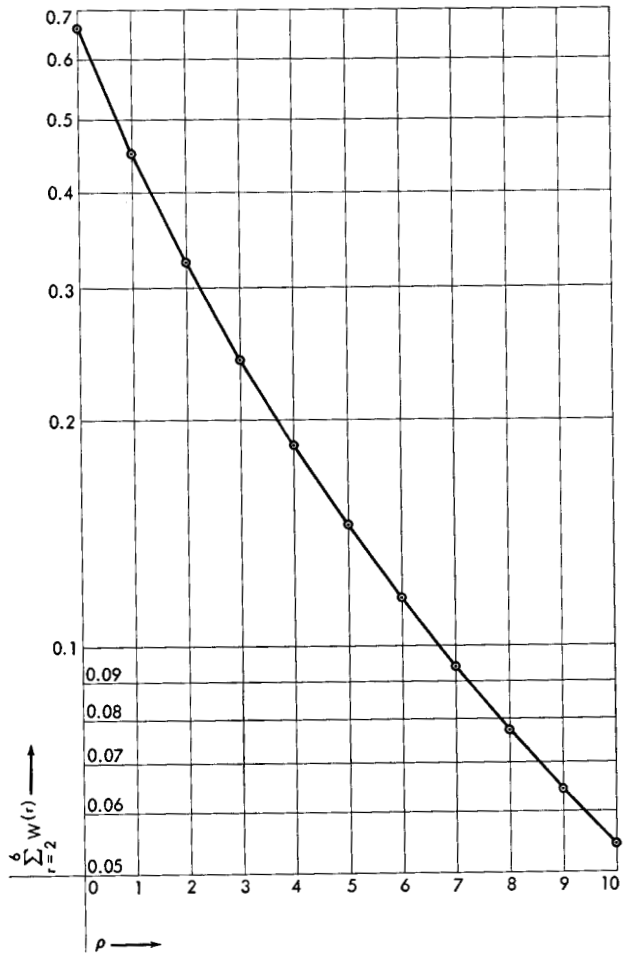


Figure 6 **Destruction of orderliness by ten shuffles.**

The total correlation per digit $\sum_{r=2}^{6} W^{(r)}$ for $\varepsilon=1/16$ *against the number* ρ *of shuffles.*

total correlation $\sum_{r=2}^{6} W^{(r)}$ decreases with ρ for $\varepsilon=1/16$ in the domain $\rho=0$ through $\rho=10$. Figure 7 shows how each $W^{(r)}(r=2, 3, 4, 5, 6)$ decreases with increasing ρ for $\varepsilon=1/16$.

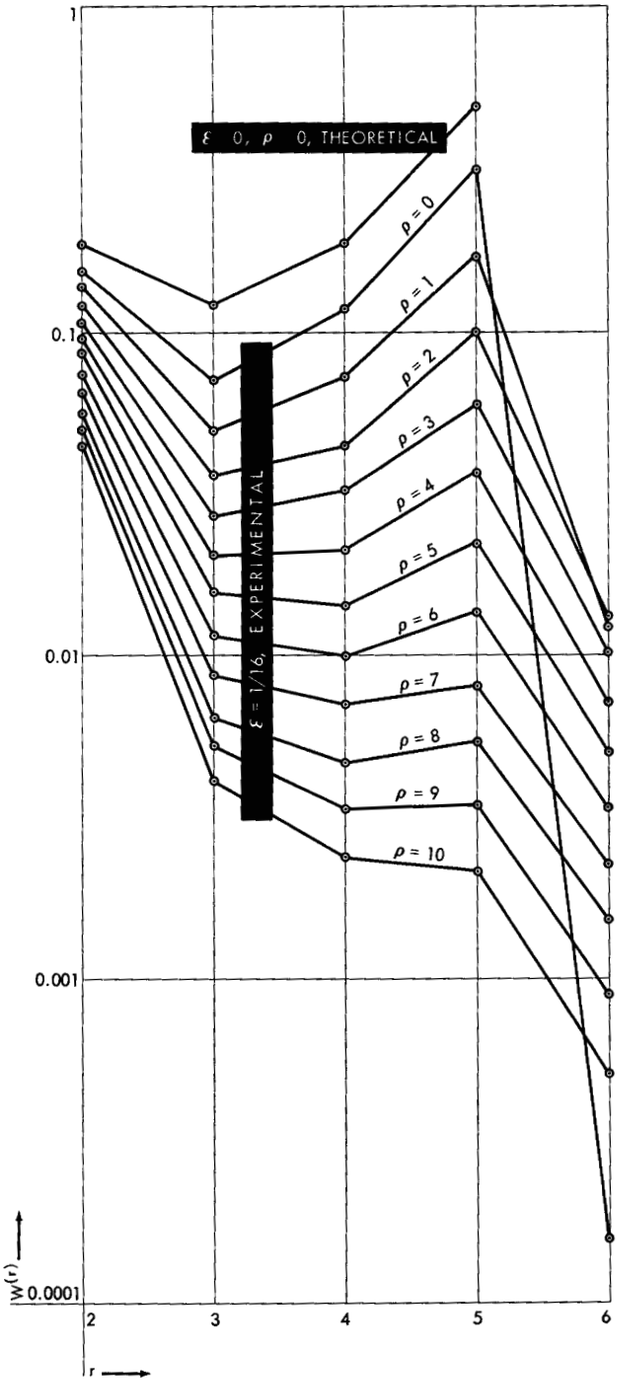
We should expect $W^{(6)}=0$ before shuffling, as stated previously, if the total length N were infinite. We must, however, expect that the experimental value of $W^{(r)}$ for a larger value of r will differ from the theoretical value (which is computed for $N\to\infty$) for the following two reasons: (1) Since the number of varieties of segments of length r is n^r, which increases rapidly with r, the number of samples of the segments in a given N-digit chain becomes more insufficient for larger r. (2) As r increases, the probability of each variety becomes smaller. But the function $-p \log p$ becomes very sensitive to a small relative change of p if p itself becomes close to zero. This will explain why we had in our experiment, for instance, $W^{(6)}=0.0001584$ instead of zero, i.e., about 10^{-39}, for $\rho=0$ and $\varepsilon=1/16$. This $W^{(6)}$, however, is very small compared with the total correlation $\sum_{r=2}^{6} W^{(r)}=0.6651$. The reason why $W^{(6)}$ with $\varepsilon=1/16$ jumps at $\rho=1$ to a much larger value, 0.01265, seems to be entirely different. Assume, for simplicity, $\varepsilon=0$. Then without shuffling, we should have $S^{(4)}=S^{(5)}=S^{(6)}=3$, which makes $W^{(6)}=-S^{(6)}+2S^{(5)}-S^{(4)}$ vanish. After a very large number of shuffles, each $S^{(r)}$ will tend to its maximum, which is r. This will make $W^{(6)}$ vanish again. At in-between stages, let us denote the increment of $S^{(r)}$ by each shuffle by $\Delta S^{(r)}$. Then, unless the condition $\Delta S^{(6)}+\Delta S^{(4)}=2\Delta S^{(5)}$ is satisfied at each stage, $W^{(6)}$ will deviate from the value zero, viz., it will become positive. Consequently, it can very well happen that $W^{(6)}$ starts from 0 at $\rho=0$ and becomes non-zero at the intermediate stages and then goes back to 0 for $\rho\to\infty$. In our experiment with $\varepsilon=1/16$, $W^{(6)}$ started from 0.0001584 at $\rho=0$, and reached a maximum, 0.01315, at $\rho=2$ and then went down to 0.0005072 at $\rho=10$. See Fig. 7.

Acknowledgment

This series of rather complicated experiments on the IBM 704 was carried out by Carol E. Shanesy and Michael Greene with the helpful advice of Robert Ramey. The author's sincere thanks are due each one of them.

Figure 7 **Decrease of orderliness as measured by correlation indices by shuffles.**

The correlation index $W^{(r)}$ against the range r, for $\varepsilon=1/16$. Each curve corresponds to a value of shuffles ρ which runs from 0 to 10. As can be expected, $W^{(r)}$ with larger r is destroyed more quickly. For comparison, the theoretical values of $W^{(r)}$ for $\varepsilon=0$ and $\rho=0$ are also plotted.



References overleaf.

References

1. L. Boltzmann, *Vorlesungen über Gastheorie,* Leipzig (1946-48).
2. L. Szilard, *Z. Physik* **53,** 840 (1929).
3. J. von Neumann, *Mathematische Grundlagen der Quantenmechanik,* Springer, Berlin (1932).
4. S. Watanabe, *Z. Physik,* **113,** 482 (1939).
5. R. V. L. Hartley, *Bell System Tech. J.* **7,** 535 (1928).
6. C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication,* University of Illinois Press, Urbana (1949).
7. N. Wiener, *Cybernetics,* John Wiley & Sons, New York (1948).
   N. Wiener, *The Extrapolation, Interpolation and Smoothing of Stationary Time Series,* John Wiley & Sons, New York (1949).
8. A. Kolmogoroff, *Bull. acad. sci. U.S.S.R. ser. math.* **5,** 3 (1942).
9. J. Rothstein, *J. Appl. Phys.* **23,** 1981 (1952) (L).
   J. Rothstein, 1954 Symposium, *IRE PGIT Transactions* **4,** p. 64.
   J. Rothstein, *Communication, Organization, and Science,* The Falcon's Wing Press, Indian Hills, Col. (1958).
10. W. J. McGill, *Psychometrika,* **19,** 97 (1954).
11. W. R. Gardner and W. J. McGill, *Psychometrika,* **21,** 219 (1956).
12. S. K. Mitra, *Contributions to the Statistical Analysis of Categorical Data,* Thesis, University of North Carolina, 1955.
13. S. Watanabe, 1954 Symposium, *IRE PGIT Transactions,* **4,** p. 85.
14. S. Watanabe, *Nuovo Cimento* Supplement (in press).

Received June 6, 1958.
Revised manuscript received July 28, 1959.