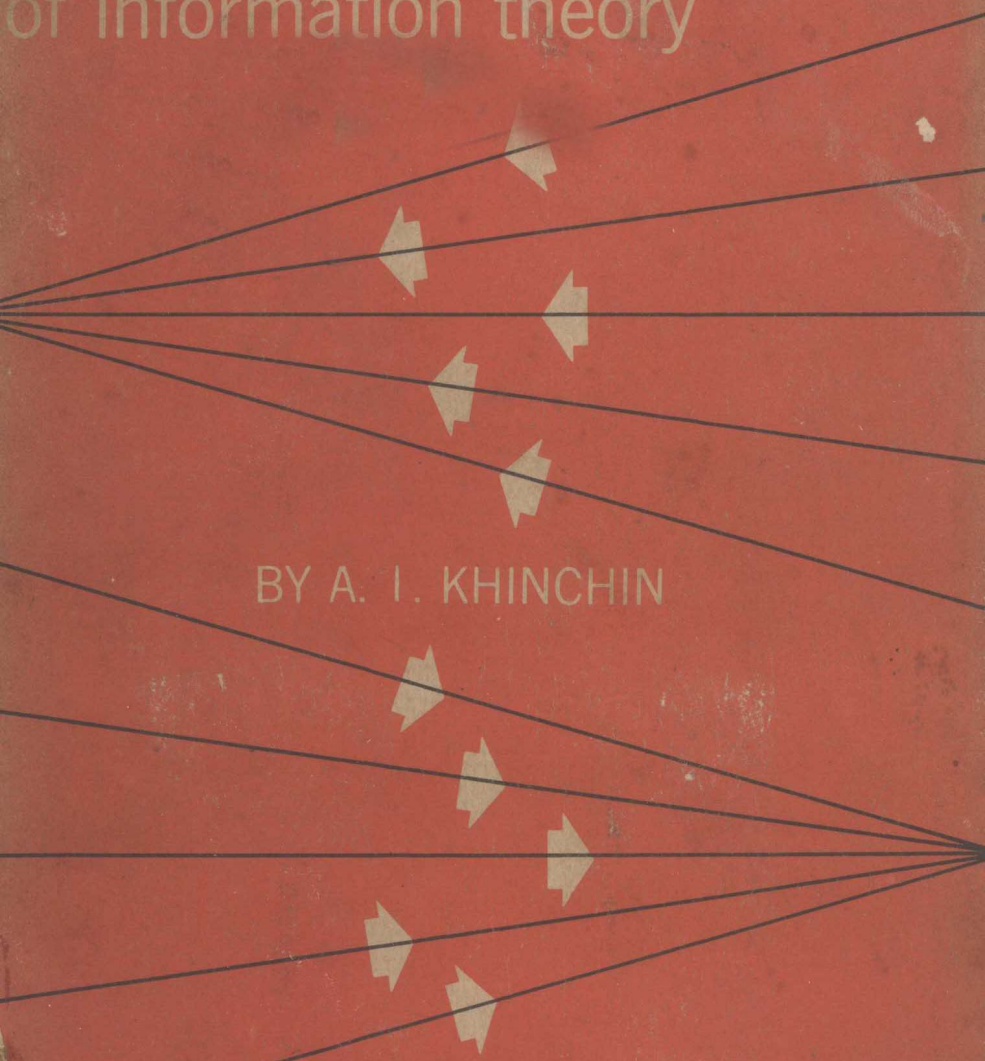
mathematical foundations of information theory



BY A. I. KHINCHIN

translated by R. A. Silverman and & M. D. Friedman

mathematical foundations of information theory

A. I. Khinchin

translated by
R. A. Silverman
institute of mathematical sciences
new york university
and
M. D. Friedman
lincoln laboratory
massachusetts institute of technology

dover publications, inc., new york

CONTENTS

The Entropy Concept in Probability Theory

On the Fundamental Theorems of Information Theory

The Entropy Concept in Probability Theory

(Uspekhi Matematicheskikh Nauk, vol. VIII, no. 3, 1953, pp. 3–20)

In his article "On the Drawing of Maps" P. L. Chebyshev beautifully expresses the nature of the relation between scientific theory and practice (discussing the case of mathematics): "The bringing together of theory and practice leads to the most favorable results; not only does practice benefit, but the sciences themselves develop under the influence of practice, which reveals new subjects for investigation and new aspects of familiar subjects." A striking example of the phenomenon described by Chebyshev is afforded by the concept of entropy in probability theory, a concept which has evolved in recent years from the needs of practice. This concept first arose in attempting to create a theoretical model for the transmission of information of various kinds. In the beginning the concept was introduced in intimate association with transmission apparatus of one kind or another; its general theoretical significance and properties, and the general nature of its application to practice were only gradually realized. As of the present, a unified exposition of the theory of entropy can be found only in specialized articles and monographs dealing with the transmission of information. Although the study of entropy has actually evolved into an important and interesting chapter of the general theory of probability, a presentation of it in this general theoretical setting has so far been lacking.

This article represents a first attempt at such a presentation. In writing it, I relied mainly on Shannon's paper "The Mathematical Theory of Communication".* However, Shannon's treatment is not always sufficiently complete and mathematically correct, so that besides having to free the theory from practical details, in many instances I have amplified and changed both the statement of definitions and the statement and proofs of theorems. There is no doubt that in the years to come the study of entropy will become a permanent part of probability theory; the work I have done seems to me to be a necessary stage in the development of this study.

§1. Entropy of Finite Schemes

In probability theory a *complete system of events* $A_1, A_2, \cdots, A_n$ means a set of events such that one and only one of them must occur at each trial (e.g., the appearance of 1, 2, 3, 4, 5, or 6 points in throwing a die). In the case $n=2$ we have a simple alternative or pair of *mutually exclusive* events (e.g., the appearance of heads or tails in tossing a coin). If we are given the events $A_1, A_2, \cdots, A_n$ of a complete system, together with their probabilities $p_1, p_2, \cdots, p_n$ $(p_i \geq 0, \sum_{i=1}^{n} p_i = 1)$, then we say that we have a *finite scheme*

$$A = \begin{pmatrix} A_1 & A_2 & \cdots & A_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}. \tag{1}$$

In the case of a "true" die, designating the appearance of i points by A_i $(1 \leq i \leq 6)$, we have the finite scheme

$$\begin{pmatrix} A_1 & A_2 & A_3 & A_4 & A_5 & A_6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix}.$$

* C. E. Shannon, Bell System Technical Journal, 27, 379-423; 623-656 (1948).

Every finite scheme describes a state of *uncertainty*. We have an experiment, the outcome of which must be one of the events $A_1, A_2, \cdots, A_n$, and we know only the probabilities of these possible outcomes. It seems obvious that the amount of uncertainty is different in different schemes. Thus, in the two simple alternatives

$$\begin{pmatrix} A_1 & A_2 \\ 0.5 & 0.5 \end{pmatrix}, \qquad \begin{pmatrix} A_1 & A_2 \\ 0.99 & 0.01 \end{pmatrix},$$

the first obviously represents much more uncertainty than the second; in the second case, the result of the experiment is "almost surely" A_1, while in the first case we naturally refrain from making any predictions. The scheme

$$\begin{pmatrix} A_1 & A_2 \\ 0.3 & 0.7 \end{pmatrix}$$

represents an amount of uncertainty intermediate between the preceding two, etc.

For many applications it seems desirable to introduce a quantity which in a reasonable way measures the amount of uncertainty associated with a given finite scheme. We shall see that the quantity

$$H(p_1, p_2, \cdots, p_n) = -\sum_{k=1}^{n} p_k \lg p_k,$$

can serve as a very suitable measure of the uncertainty of the finite scheme (1); the logarithms are taken to an arbitrary but fixed base, and we always take $p_k \lg p_k = 0$ if $p_k = 0$. We shall call the quantity $H(p_1, p_2, \cdots, p_n)$ the *entropy* of the finite scheme (1), pursuing a physical analogy which there is no need to go into here. We now convince ourselves that this function actually has a number of properties which we might expect,

of a reasonable measure of uncertainty of a finite scheme.

First of all, we see immediately that $H(p_1, p_2, \cdots, p_n)=0$, if and only if one of the numbers $p_1, p_2, \cdots, p_n$ is one and all the others are zero. But this is just the case where the result of the experiment can be predicted beforehand with complete certainty, so that there is no uncertainty as to its outcome. In all other cases the entropy is positive.

Furthermore, for fixed n it is obvious that the scheme with the most uncertainty is the one with equally likely outcomes, i.e., $p_k=1/n$ $(k=1, 2, \cdots, n)$, and in fact the entropy assumes its largest value for just these values of the variables p_k. The easiest way to see this is to use an inequality which is valid for any continuous convex function $\varphi(x)$

$$\varphi\left(\frac{1}{n}\sum_{k=1}^{n} a_k\right) \leq \frac{1}{n}\sum_{k=1}^{n} \varphi(a_k),$$

where $a_1, a_2, \cdots, a_n$ are any positive numbers. Setting $a_k=p_k$ and $\varphi(x)=x \lg x$, and bearing in mind that $\sum_{k=1}^{n} p_k=1$, we find

$$\varphi\left(\frac{1}{n}\right)=\frac{1}{n}\lg\frac{1}{n} \leq \frac{1}{n}\sum_{k=1}^{n} p_k \lg p_k=-\frac{1}{n}H(p_1, p_2, \cdots, p_n),$$

whence

$$H(p_1, p_2, \cdots, p_n) \leq \lg n=H\left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right), \quad \text{Q.E.D.}$$

Suppose now we have two finite schemes

$$A=\begin{pmatrix} A_1 & A_2 \cdots A_n \\ p_1 & p_2 \cdots p_n \end{pmatrix}, \quad B=\begin{pmatrix} B_1 & B_2 \cdots B_m \\ q_1 & q_2 \cdots q_m \end{pmatrix},$$

and let these two schemes be (mutually) independent, i.e., the probability π_{kl} of the joint occurrence of the events A_k and B_l is $p_k q_l$. Then, the set of events $A_k B_l$ $(1 \leq k \leq n,\ 1 \leq l \leq m)$,

with probabilities π_{kl} represents another finite scheme, which we call the *product* of the schemes A and B and designate by AB. Let $H(A)$, $H(B)$, and $H(AB)$ be the corresponding entropies of the schemes A, B, and AB. Then

$$H(AB)=H(A)+H(B), \tag{2}$$

for, in fact

$$-H(AB)=\sum_k\sum_l \pi_{kl}\lg\pi_{kl}=\sum_k\sum_l p_kq_l(\lg p_k+\lg q_l)=$$
$$=\sum_k p_k\lg p_k\sum_l q_l+\sum_l q_l\lg q_l\sum_k p_k=-H(A)-H(B).$$

We now turn to the case where the schemes A and B are (mutually) dependent. We denote by q_{kl} the probability that the event B_l of the scheme B occurs, given that the event A_k of the scheme A occurred, so that

$$\pi_{kl}=p_kq_{kl}\ (1\leq k\leq n,\ 1\leq l\leq m).$$

Then

$$-H(AB)=\sum_k\sum_l p_kq_{kl}(\lg p_k+\lg q_{kl})=$$
$$=\sum_k p_k\lg p_k\sum_l q_{kl}+\sum_k p_k\sum_l q_{kl}\lg q_{kl}.$$

Here $\sum_l q_{kl}=1$ for any k, and the sum $-\sum_l q_{kl}\lg q_{kl}$ can be regarded as the conditional entropy $H_k(B)$ of the scheme B, calculated on the assumption that the event A_k of the scheme A occurred. We obtain

$$H(AB)=H(A)+\sum_k p_kH_k(B).$$

The conditional entropy $H_k(B)$ is obviously a random variable in the scheme A; its value is completely determined by the knowledge of which event A_k of the scheme A actually occurred. Therefore, the last term of the right side is the *mathematical*

expectation of the quantity $H(B)$ *in the scheme* A, which we shall designate by $H_A(B)$. Thus in the most general case, we have

$$H(AB)=H(A)+H_A(B). \tag{3}$$

It is self-evident that the relation (3) reduces to (2) in the special case where the schemes A and B are independent.

It is also interesting to note that in all cases $H_A(B)\leq H(B)$. It is reasonable to interpret this inequality as saying that, on the average, knowledge of the outcome of the scheme A can only decrease the uncertainty of the scheme B. To prove this, we observe that any continuous convex function $f(x)$ obeys the inequality*

$$\sum_k \lambda_k f(x_k)\geq f(\sum_k \lambda_k x_k),$$

if $\lambda_k\geq 0$ and $\sum_k \lambda_k=1$. Therefore, setting $f(x)=x\lg x$, $\lambda_k=p_k$, $x_k=q_{kl}$, we find for arbitrary l that

$$\sum_k p_k q_{kl}\lg q_{kl}\geq(\sum_k p_k q_{kl})\lg(\sum_k p_k q_{kl})=q_l\lg q_l,$$

since obviously $\sum_k p_k q_{kl}=q_l$. Summing over l, we obtain on the left side the quantity

$$\sum_k p_k\sum_l q_{kl}\lg q_{kl}=-\sum_k p_k H_k(B)=-H_A(B),$$

and consequently we find

$$-H_A(B)\geq\sum_l q_l\lg q_l=-H(B), \quad \text{Q.E.D.}$$

If we carry out an experiment the possible outcomes of which are described by the given scheme A, then in doing so we obtain some *information* (i.e., we find out which of the events

* See, for example, Hardy, Littlewood, and Pólya, *Inequalities*, Cambridge University Press, 1934.

A_k actually occurs), and the uncertainty of the scheme is completely eliminated. Thus, we can say that the information given us by carrying out some experiment consists in removing the uncertainty which existed before the experiment. The larger this uncertainty, the larger we consider to be the amount of information obtained by removing it. Since we agreed to measure the uncertainty of a finite scheme A by its entropy $H(A)$, it is natural to express the amount of information given by removing this uncertainty by an increasing function of the quantity $H(A)$. The choice of this function means the choice of some unit for the quantity of information and is therefore fundamentally a matter of indifference. However, the properties of entropy which we demonstrated above show that it is especially convenient to take this quantity of information proportional to the entropy. Indeed, consider two finite schemes A and B and their product AB. Realization of the scheme AB is obviously equivalent to realization of both of the schemes A and B. Therefore, if the two schemes A and B are independent, it is natural to require the information given by the realization of the scheme AB to be the sum of the two amounts of information given by the realization of the schemes A and B; since in this case

$$H(AB)=H(A)+H(B),$$

this requirement will actually be met, if we consider the amount of information given by the realization of a finite scheme to be proportional to the entropy of the scheme. Of course, the constant of proportionality can be taken as unity, since this choice corresponds merely to a choice of units. Thus, in all that follows, we can consider the amount of information given

by the realization of a finite scheme to be equal to the entropy of the scheme. This stipulation makes the concept of entropy especially significant for information theory.

In view of this stipulation, let us consider the case of two dependent schemes A and B and the corresponding relation (3). The amount of information given by the realization of the scheme AB is equal to $H(AB)$. However, as explained above, in the general case, this cannot be equal to $H(A)+H(B)$. Indeed, consider the extreme case where knowledge of the outcome of the scheme A also determines with certainty the outcome of the scheme B, so that each event A_k of the scheme A can occur only in conjunction with a specific event B_l of the scheme B. Then, after realization of the scheme A, the scheme B completely loses its uncertainty, and we have $H_A(B)=0$; moreover, in this case realization of the scheme B obviously gives no further information, and we have $H(AB)=H(A)$, so that relation (3) is indeed satisfied. In all cases, the quantity $H_k(B)$ introduced above is the amount of information given by the scheme B, given that the event A_k occurred in the scheme A; therefore the quantity $H_A(B)=\sum_k p_k H_k(B)$ is the mathematical expectation of the amount of additional information given by realization of the scheme B after realization of scheme A and reception of the corresponding information. Therefore, the relation (3) has the following very reasonable interpretation: *The amount of information given by the realization of the two finite schemes A and B, equals the amount of information given by the realization of scheme A, plus the mathematical expectation of the amount of additional information given by the realization of scheme B after the realization of the scheme A.* In just the same way we can give an

entirely reasonable interpretation of the general inequality $H_A(B) \leq H(B)$ proved above: *The amount of information given by the realization of a scheme B can only decrease if another scheme A is realized beforehand.*

§2. The Uniqueness Theorem

Among the properties of entropy which we have proved, we can consider the following two as basic:

1. For given n and for $\sum_{k=1}^{n} p_k = 1$, the function $H(p_1, p_2, \cdots, p_n)$ takes its largest value for $p_k = \frac{1}{n}$ $(k=1, 2, \cdots, n)$.

2. $H(AB) = H(A) + H_A(B)$.

We add to these two properties a third, which obviously must be satisfied by any reasonable definition of entropy. Since the schemes

$$\begin{pmatrix} A_1 & A_2 & \cdots & A_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix} \text{ and } \begin{pmatrix} A_1 & A_2 & \cdots & A_n & A_{n+1} \\ p_1 & p_2 & \cdots & p_n & 0 \end{pmatrix},$$

are obviously not substantively different, we must have

3. $H(p_1, p_2, \cdots, p_n, 0) = H(p_1, p_2, \cdots, p_n)$. (Adding the impossible event or any number of impossible events to a scheme does not change its entropy.) We now prove the following important proposition:

Theorem 1.

Let $H(p_1, p_2, \cdots, p_n)$ be a function defined for any integer n and for all values $p_1, p_2, \cdots, p_n$ such that $p_k \geq 0$ $(k=1,2,\cdots, n)$, $\sum_{k=1}^{n} p_k = 1$. If for any n this function is continuous with respect to all its arguments, and if it has the properties 1, 2, *and* 3, *then*

$$H(p_1, p_2, \cdots, p_n) = -\lambda \sum_{k=1}^{n} p_k \lg p_k,$$

where λ is a positive constant.

This theorem shows that the expression for the entropy of a finite scheme which we have chosen is the only one possible if we want it to have certain general properties which seem necessary in view of the actual meaning of the concept of entropy (as a measure of uncertainty or as an amount of information).

Proof.

For brevity we set

$$H\left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right)=L(n);$$

we shall show that $L(n)=\lambda \lg n$, where λ is a positive constant. By 3 and 1, we have

$$L(n)=H\left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}, 0\right) \leq H\left(\frac{1}{n+1}, \frac{1}{n+1}, \cdots, \frac{1}{n+1}\right)=L(n+1),$$

so that $L(n)$ is a non-decreasing function of n. Let m and r be positive integers. Consider m mutually independent finite schemes $S_1, S_2, \cdots, S_m$, each of which contains r equally likely events, so that

$$H(S_k)=H\left(\frac{1}{r}, \frac{1}{r}, \cdots, \frac{1}{r}\right)=L(r) \quad (1 \leq k \leq m).$$

By Property 2 (generalized to the case of m schemes) we have, in view of the independence of the schemes S_k

$$H(S_1 S_2 \cdots S_m)=\sum_{k=1}^{m} H(S_k)=mL(r).$$

But the product scheme $S_1 S_2 \cdots S_m$ obviously consists of r^m equally likely events, so that its entropy is $L(r^m)$. Therefore we have

$$L(r^m)=mL(r), \tag{4}$$

and similarly, for any other pair of positive integers n and s

$$L(s^n)=nL(s). \tag{5}$$

Now let the numbers r, s, and n be given arbitrarily, but let the number m be determined by the inequalities

$$r^m \leq s^n < r^{m+1}, \tag{6}$$

whence

$$\begin{gathered} m \lg r \leq n \lg s < (m+1) \lg r, \\ \frac{m}{n} \leq \frac{\lg s}{\lg r} < \frac{m}{n} + \frac{1}{n}. \end{gathered} \tag{7}$$

It follows from (6) by the monotonicity of the function $L(n)$ that

$$L(r^m) \leq L(s^n) \leq L(r^{m+1}),$$

and, consequently, by (4) and (5)

$$mL(r) \leq nL(s) \leq (m+1)L(r),$$

so that

$$\frac{m}{n} \leq \frac{L(s)}{L(r)} \leq \frac{m}{n} + \frac{1}{n}. \tag{8}$$

Finally, it follows from (7) and (8) that

$$\left| \frac{L(s)}{L(r)} - \frac{\lg s}{\lg r} \right| \leq \frac{1}{n}$$

Since the left side of this inequality is independent of m, and since n can be chosen arbitrarily large in the right side

$$\frac{L(s)}{\lg s} = \frac{L(r)}{\lg r},$$

which, in view of the arbitrariness of r and s, means that

$$L(n)=\lambda \lg n,$$

where λ is a constant. By the monotonicity of the function $L(n)$, we have $\lambda \geqq 0$, and our assertion is proved.

This assertion represents the special case $p_k=1/n$ $(1 \leqq k \leqq n)$ of the theorem to be proved. We now consider the more general case, where the p_k $(k=1,2,\cdots,n)$ are any rational numbers. Let

$$p_k=\frac{g_k}{g} \quad (k=1,2,\cdots,n),$$

where all the g_k are positive integers and $\sum_{k=1}^{n} g_k=g$. Let the finite scheme A consist of n events with probabilities $p_1, p_2,\cdots, p_n$. Our problem consists in defining the entropy of this scheme. To this end, we consider a second scheme B, which is dependent on A and is defined as follows: The scheme B contains g events $B_1, B_2,\cdots, B_g$, which we devide into n groups, containing $g_1, g_2,\cdots, g_n$ events, respectively. If the event A_k occurred in scheme A, then in scheme B all the g_k events of the k'th group have the same probability $1/g_k$, and all the events of the other groups have probability zero (are impossible). Thus, given any outcome A_k of the scheme A, the scheme B reduces to a system of g_k equally likely events, so that the conditional entropy

$$H_k(B)=H(1/g_k, 1/g_k,\cdots, 1/g_k)=L(g_k)=\lambda \lg g_k,$$

which means that

$$H_A(B)=\sum_{k=1}^{n} p_k H_k(B)=\lambda \sum_{k=1}^{n} p_k \lg g_k=\lambda \sum_{k=1}^{n} p_k \lg p_k+\lambda \lg g. \qquad (9)$$

We return now to the product scheme AB, consisting of the

events A_kB_l $(1 \leq k \leq n,\ 1 \leq l \leq g)$. Such an event is possible only if B_l belongs to the k'th group. Thus, the number of possible events A_kB_l for a given k is g_k, and the total number of possible events in the scheme AB is $\sum_{k=1}^{n} g_k = g$. The probability of each possible event A_kB_l is obviously $p_k/g_k = 1/g$, i.e., is the same for all the events. Thus, the scheme AB consists of g equally likely events, and therefore

$$H(AB) = L(g) = \lambda \lg g.$$

Using property (2) and relation (9), we find

$$\lambda \lg g \leq H(A) + \lambda \sum_{k=1}^{n} p_k \lg p_k + \lambda \lg g,$$

whence

$$H(A) \geq H(p_1, p_2, \cdots, p_n) = -\lambda \sum_{k=1}^{n} p_k \lg p_k. \tag{10}$$

Finally, relation (10) which we have proved for rational $p_1, p_2, \cdots, p_n$, must be valid for any values of its arguments because of the postulated continuity of the function $H(p_1, p_2, \cdots, p_n)$. Thus the proof of Theorem 1 is complete.

§3. Entropy of Markov chains

Suppose we have a simple stationary Markov chain with a finite number of states $A_1, A_2, \cdots, A_n$ and with the transition probability matrix p_{ik} $(i, k = 1,2, \cdots, n)$. We denote by P_k the probability of the state A_k $(1 \leq k \leq n)$, so that in particular

$$\sum_{k=1}^{n} P_k p_{kl} = P_l \quad (l = 1,2, \cdots, n). \tag{11}$$

If the system is in state A_i, then its transitions to the different states A_k $(k = 1,2, \cdots, n)$ form a finite scheme

$$\begin{pmatrix} A_1 & A_2 & \cdots & A_n \\ p_{i1} & p_{i2} & \cdots & p_{in} \end{pmatrix},$$

the entropy of which

$$H_i = -\sum_{k=1}^{n} p_{ik} \lg p_{ik}$$

depends on i and can be regarded as *a measure of the amount of information obtained when the Markov chain moves one step ahead, starting from the initial state* A_i. The average of this quantity over all initial states, i.e., the quantity

$$H = \sum_{i=1}^{n} P_i H_i = -\sum_{i=1}^{n} \sum_{k=1}^{n} P_i p_{ik} \lg p_{ik},$$

is therefore to be regarded as *a measure of the average amount of information obtained when the given Markov chain moves one step ahead.* This quantity H, which we shall call the *entropy* of the chain in question obviously characterizes the chain as a whole; it is clear that it is uniquely determined by giving the state probabilities P_i and the transition probabilities p_{ik} $(1 \leq i \leq n,\ 1 \leq k \leq n)$.

All the concepts which are defined for moving one step ahead can be easily and naturally generalized to the case of moving ahead an arbitrary number of steps r. If the system is in state A_i, then it is easy to calculate the probability that in the next r trials we shall find it in the states $A_{k_1}, A_{k_2}, \cdots, A_{k_r}$ in turn, where $k_1, k_2, \cdots, k_r$ are arbitrary numbers from 1 to n. Thus, the subsequent fate of a system initially in the state A_i in the next r trials is described by a finite scheme (with n^r events), with a definite entropy which we designate by $H_i^{(r)}$ and regard as a measure of the amount of information obtained in moving ahead r steps in the chain, starting from the initial state A_i. The quantity

$$H^{(r)} = \sum_{i=1}^{n} P_i H_i^{(r)}$$