Chapter 14
Information and Entropy

Abstract Section 14.1 presents the definitions and key properties of information and entropy. Section 14.2 discusses the entropy of a (stationary) finite Markov chain. The Law of Large Numbers is proved for the amount of information contained in a message that is a long sequence of successive states of a Markov chain, and the asymptotic behaviour of the number of the most common states in a sequence of successive values of the chain is established. Applications of this result to coding are discussed.

14.1 The Definitions and Properties of Information and Entropy

Suppose one conducts an experiment whose outcome is not predetermined. The term "experiment" will have a broad meaning. It may be a test of a new device, a satellite launch, a football match, a referendum and so on. If, in a football match, the first team is stronger than the second, then the occurrence of the event A that the first team won carries little significant information. On the contrary, the occurrence of the complementary event $\overline{A}$ contains a lot of information. The event B that a leading player of the first team was injured does contain information concerning the event A. But if it was the first team's doctor who was injured then that would hardly affect the match outcome, so such an event B carries no significant information about the event A.

The following quantitative measure of information is conventionally adopted. Let A and B be events from some probability space $\langle \Omega, \mathfrak{F}, \mathbf{P} \rangle$.

Definition 14.1.1 The *amount of information about the event A contained in the event (message) B* is the quantity

$$I(A|B) := \log \frac{\mathbf{P}(A|B)}{\mathbf{P}(A)}.$$

The notions of the "amount of information" and "entropy" were introduced by C.E. Shannon in 1948. For some special situations the notion of amount of information had also been considered in earlier papers (e.g., by R.V.L. Hartley, 1928). The exposition in Sect. 14.2 of this chapter is substantially based on the paper of A.Ya. Khinchin [21].

A.A. Borovkov, *Probability Theory*, Universitext,
DOI 10.1007/978-1-4471-5201-9_14, © Springer-Verlag London 2013

The occurrence of the event $B = A$ may be interpreted as the message that A took place.

Definition 14.1.2 The number $I(A) := I(A|A)$ is called the *amount of information contained in the message* A:

$$I(A) := I(A|A) = -\log \mathbf{P}(A).$$

We see from this definition that the larger the probability of the event A, the smaller $I(A)$. As a rule, the logarithm to the base 2 is used in the definition of information. Thus, say, the message that a boy (or girl) was born in a family carries a unit of information (it is supposed that these events are equiprobable, and $-\log_2 p = 1$ for $p = 1/2$). Throughout this chapter, we will write just $\log x$ for $\log_2 x$.

If the events A and B are independent, then $I(A|B) = 0$. This means that the event B does not carry any information about A, and vice versa. It is worth noting that we always have

$$I(A|B) = I(B|A).$$

It is easy to see that if the events A and B are independent, then

$$I(AB) = I(A) + I(B). \tag{14.1.1}$$

Consider an example. Let a chessman be placed at random on one of the squares of a chessboard. The information that the chessman is on square number k (the event A) is equal to $I(A) = \log 64 = 6$. Let B_1 be the event that the chessman is in the i-th row, and B_2 that the chessman is in the j-th column. The message A can be transmitted by transmitting B_1 first and then B_2. We have

$$I(B_1) = \log 8 = 3 = I(B_2).$$

Therefore

$$I(B_1) + I(B_2) = 6 = I(A),$$

so that transmitting the message A "by parts" requires communicating the same amount of information (which is equal to 6) as transmitting A itself. One could give other examples showing that the introduced numerical characteristics are quite natural.

Let G be an experiment with outcomes $E_1, \dots, E_N$ occurring with probabilities $p_1, \dots, p_N$.

The information resulting from the experiment G is a random variable $J_G = J_G(\omega)$ assuming the value $-\log p_j$ on the set E_j, $j = 1, \dots, N$.

Thus, if in the probability space $\langle \Omega, \mathfrak{F}, \mathbf{P} \rangle$ corresponding to the experiment G, Ω coincides with the set $(E_1, \dots, E_N)$, then $J_G(\omega) = I(\omega)$.

Definition 14.1.3 The expectation of the information obtained in the experiment G, $\mathbf{E}J_G = -\sum p_j \log p_j$, is called the *entropy* of the experiment. We shall denote it by

$$H_{\mathbf{p}} = H(G) := -\sum_{j=1}^{N} p_j \log p_j,$$

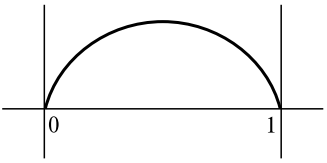

Fig. 14.1 The plot of the entropy $f(p)$ of a random experiment with two outcomes

where $\mathbf{p} = (p_1, \ldots, p_N)$. For $p_j = 0$, by continuity we set $p_j \log p_j$ to be equal to zero.

The entropy of an experiment is, in a sense, a measure of its uncertainty. Let, for example, our experiment have two outcomes A and B with probabilities p and $q = 1 - p$, respectively. The entropy of the experiment is equal to

$$H_{\mathbf{p}} = -p \log p - (1 - p) \log(1 - p) = f(p).$$

The graph of this function is depicted in Fig. 14.1.

The only maximum of $f(p)$ equals $\log 2 = 1$ and is attained at the point $p = 1/2$. This is the case of maximum uncertainty. If p decreases, then the uncertainty also decreases together with $H_{\mathbf{p}}$, and $H_{\mathbf{p}} = 0$ for $\mathbf{p} = (0, 1)$ or $(0, 1)$.

The same properties can easily be seen in the general case as well.

The properties of entropy.

1. *$H(G) = 0$ if and only if there exists a j, $1 \le j \le N$, such that $p_j = \mathbf{P}(E_j) = 1$.*
2. *$H(G)$ attains its maximum when $p_j = 1/N$ for all j.*

Proof The second derivative of the function $\beta(x) = x \log x$ is positive on $[0, 1]$, so that $\beta(x)$ is convex. Therefore, for any $q_i \ge 0$ such that $\sum_{i=1}^{N} q_i = 1$, and any $x_i \ge 0$, one has the inequality

$$\beta\left(\sum_{i=1}^{N} q_i x_i\right) \le \sum_{i=1}^{N} q_i \beta(x_i).$$

If we take $q_i = 1/N$, $x_i = p_i$, then

$$\left(\frac{1}{N}\sum_{i=1}^{N} p_i\right) \log\left(\frac{1}{N}\sum_{i=1}^{N} p_i\right) \le \sum_{i=1}^{N} \frac{1}{N} p_i \log p_i.$$

Setting $\mathbf{u} := (\frac{1}{N}, \ldots, \frac{1}{N})$ we obtain from this that

$$-\log \frac{1}{N} = \log N = H_{\mathbf{u}} \ge -\sum_{i=1}^{N} p_i \log p_i = H_{\mathbf{p}}. \qquad \square$$

Note that if the entropy $H(G)$ equals its maximum value $H(G) = \log N$, then $J_G(\omega) = \log N$ with probability 1, i.e. the information $J_G(\omega)$ becomes constant.

3. Let G_1 and G_2 be two independent experiments. We write down the outcomes and their probabilities in these experiments in the following way:

$$G_1 = \begin{pmatrix} E_1, \dots, E_N \\ p_1, \dots, p_N \end{pmatrix}, \qquad G_2 = \begin{pmatrix} A_1, \dots, A_M \\ q_1, \dots, q_M \end{pmatrix}.$$

Combining the outcomes of these two experiments we obtain a new experiment

$$G = G_1 \times G_2 = \begin{pmatrix} E_1A_1, E_1A_2, \dots, E_NA_M \\ p_1q_1, p_1q_2, \dots, p_Nq_M \end{pmatrix}.$$

The information J_G obtained as a result of this experiment is a random variable taking values $-\log p_i q_j$ with probabilities $p_i q_j$, $i = 1, \dots, N$; $j = 1, \dots, M$. But the sum $J_{G_1} + J_{G_2}$ of two independent random variables equal to the amounts of information obtained in the experiments G_1 and G_2, respectively, clearly has the same distribution. Thus the *information obtained in a sequence of independent experiments is equal to the sum of the information from these experiments*. Since in that case clearly

$$\mathbf{E}J_G = \mathbf{E}J_{G_1} + \mathbf{E}J_{G_2},$$

we have that *for independent G_1 and G_2 the entropy of the experiment G is equal to the sum of the entropies of the experiments G_1 and G_2*:

$$H(G) = H(G_1) + H(G_2).$$

4. If the experiments G_1 and G_2 are *dependent*, then the experiment G can be represented as

$$G = \begin{pmatrix} E_1A_1, E_1A_2, \dots, E_NA_M \\ q_{11}, q_{12}, \dots, q_{NM} \end{pmatrix}$$

with $q_{ij} = p_i p_{ij}$, where p_{ij} is the conditional probability of the event A_j given E_i, so that

$$\sum_{j=1}^{M} q_{ij} = p_i = \mathbf{P}(E_i), \quad i = 1, \dots, N;$$

$$\sum_{j=1}^{N} q_{ij} = q_j = \mathbf{P}(A_i), \quad j = 1, \dots, M.$$

In this case the equality $J_G = J_{G_1} + J_{G_2}$, generally speaking, does not hold. Introduce a random variable J_2^* which is equal to $-\log p_{ij}$ on the set E_iA_j. Then evidently $J_G = J_{G_1} + J_2^*$. Since

$$\mathbf{P}(A|E_i) = p_{ij},$$

the quantity J_2^* for a fixed i can be considered as the information from the experiment G_2 given the event E_i occurred. We will call the quantity

$$\mathbf{E}\big(J_2^*|E_i\big) = -\sum_{j=1}^{M} p_{ij} \log p_{ij}$$

the *conditional entropy* $H(G_2|E_1)$ *of the experiment* G_2 *given* E_i, and the quantity

$$\mathbf{E}J_2^* = -\sum_{i,j} q_{ij} \log p_{ij} = \sum_i p_i H(G_2|E_1)$$

the *conditional entropy* $H(G_2|G_1)$ *of the experiment* G_2 *given* G_1. In this notation, we obviously have

$$H(G) = H(G_1) + H(G_2|G_1).$$

We will prove that in this equality we always have

$$H(G_2|G_1) \le H(G_2),$$

i.e. *for two experiments* G_1 *and* G_2 *the entropy* $H(G)$ *never exceeds the sum of the entropies* $H(G_1)$ *and* $H(G_2)$:

$$H(G) = H(G_1 \times G_2) \le H(G_1) + H(G_2).$$

Equality takes place here only when $q_{ij} = p_i q_j$, *i.e. when* G_1 *and* G_2 *are independent.*

Proof First note that, for any two distributions $(u_1, \dots, u_n)$ and $(v_1, \dots, v_n)$, one has the inequality

$$-\sum_i u_i \log u_i \le -\sum_i u_i \log v_i, \tag{14.1.2}$$

equality being possible here only if $v_i = u_i$, $i = 1, \dots, n$. This follows from the concavity of the function $\log x$, since it implies that, for any $a_i > 0$,

$$\sum_i u_i \log a_i \le \log\left(\sum_i u_i a_i\right),$$

equality being possible only if $a_1 = a_2 = \cdots = a_n$. Putting $a_i = v_i/u_i$, we obtain relation (14.1.2).

Next we have

$$H(G_1) + H(G_2) = -\sum_{i,j} q_{ij}(\log p_i + \log q_j) = -\sum_{i,j} q_{ij} \log p_i q_j,$$

and because $\{p_i q_j\}$ is obviously a distribution, by virtue of (14.1.2)

$$-\sum q_{ij} \log p_i q_j \ge -\sum q_{ij} \log q_{ij} = H(G)$$

holds, and equality is possible here only if $q_{ij} = p_i q_j$. □

5. As we saw when considering property 3, the information obtained as a result of the experiment G_1^n consisting of n independent repetitions of the experiment G_1 is equal to

$$J_{G_1^n} = -\sum_{j=1}^{N} \nu_j \log p_j,$$

where ν_j is the number of occurrences of the outcome E_j. By the law of large numbers, $\nu_j/n \xrightarrow{p} p_j$ as $n \to \infty$, and hence

$$\frac{1}{n} J_{G_1^n} \xrightarrow{p} H(G_1) = H_p.$$

To conclude this section, we note that the measure of the amount of information resulting from an experiment we considered here can be *derived* as the only possible one (up to a constant multiplier) if one starts with a few simple requirements that are natural to impose on such a quantity.[1]

It is also interesting to note the connections between the above-introduced notions and large deviation probabilities. As one can see from Theorems 5.1.2 and 5.2.4, the difference between the "biased" entropy $-\sum p_j^* \ln p_j$ and the entropy $-\sum p_j^* \ln p_j^*$ ($p_j^* = \nu_j/n$ are the relative frequencies of the outcomes E_j) is an analogue of the deviation function (see Sect. 8.8) in the multi-dimensional case.

14.2 The Entropy of a Finite Markov Chain. A Theorem on the Asymptotic Behaviour of the Information Contained in a Long Message; Its Applications

14.2.1 The Entropy of a Sequence of Trials Forming a Stationary Markov Chain

Let $\{X_k\}_{k=1}^{\infty}$ be a stationary finite Markov chain with one class of essential states without subclasses, $E_1, \ldots, E_N$ being its states. Stationarity of the chain means that $\mathbf{P}(X_1 = j) = \pi_j$ coincide with the stationary probabilities. It is clear that

$$\mathbf{P}(X_2 = j) = \sum_k \pi_k p_{kj} = \pi_j, \qquad \mathbf{P}(X_3 = j) = \pi_j, \quad \text{and so on.}$$

Let G_k be an experiment determining the value of X_k (i.e. the state the system entered on the k-th step). If $X_{k-1} = i$, then the entropy of the k-th step equals

$$H(G_k | X_{k-1} = i) = -\sum_j p_{ij} \log p_{ij}.$$

By definition, the entropy of a stationary Markov chain is equal to

$$H = \mathbf{E} H(G_k | X_{k-1}) = H(G_k | G_{k-1}) = -\sum_i \pi_i \sum_j p_{ij} \log p_{ij}.$$

Consider the first n steps $X_1, \ldots, X_n$ of the Markov chain. By the Markov property, the entropy of this composite experiment $G^{(n)} = G_1 \times \cdots \times G_n$ is equal to

[1] See, e.g., [11].

$$H\left(G^{(n)}\right) = H(G_1) + H(G_2|G_1) + \cdots + H(G_n|G_{n-1})$$
$$= -\sum \pi_j \log \pi_j + (n-1)H \sim nH$$

as $n \to \infty$. If X_k were independent then, as we saw, we would have exact equality here.

14.2.2 The Law of Large Numbers for the Amount of Information Contained in a Message

Now consider a finite sequence $(X_1, \ldots, X_n)$ as a message (event) C_n and denote, as before, by $I(C_n) = -\log \mathbf{P}(C_n)$ the amount of information contained in C_n. The value of $I(C_n)$ is a function on the space of elementary outcomes equal to the information $J_{G^{(n)}}$ contained in the experiment $G^{(n)}$. We now show that, with probability close to 1, this information behaves asymptotically as nH, as was the case for independent X_k. Therefore H is essentially the average information per trial in the sequence $\{X_k\}_{k=1}^{\infty}$.

Theorem 14.2.1 *As $n \to \infty$,*

$$\frac{I(C_n)}{n} = \frac{-\log \mathbf{P}(C_n)}{n} \xrightarrow{a.s.} H.$$

This means that, for any $\delta > 0$, the set of all messages C_n can be decomposed into two classes. For the first class, $|I(C_n)/n - H| < \delta$, and the sum of the probabilities of the elements of the second class tends to 0 as $n \to \infty$.

Proof Construct from the given Markov chain a new one $\{Y_k\}_{k=1}^{\infty}$ by setting $Y_k := (X_k, X_{k+1})$. The states of the new chain are pairs of states (E_i, E_j) of the chain $\{X_k\}$ with $p_{ij} > 0$. The transition probabilities are obviously given by

$$p_{(i,j)(k,l)} = \begin{cases} 0, & j \neq k, \\ p_{kl}, & j = k. \end{cases}$$

Note that one can easily prove by induction that

$$p_{(i,j)(k,l)}(n) = p_{jk}(n-1)\, p_{kl}. \tag{14.2.1}$$

From the definition of $\{Y_k\}$ it follows that the ergodic theorem holds for this chain. This can also be seen directly from (14.2.1), the stationary probabilities being

$$\lim_{n\to\infty} p_{(i,j)(k,l)}(n) = \pi_k p_{kl}.$$

Now we will need the law of large numbers for the number of visits $m_{(k,l)}(n)$ of the chain $\{Y_k\}_{k=1}^{\infty}$ to state (k,l) over time n. By virtue of this law (see Theorem 13.4.4),

$$\frac{m_{(k,l)}(n)}{n} \xrightarrow{a.s.} \pi_k p_{kl} \quad \text{as } n \to \infty.$$

Consider the random variable $\mathbf{P}(C_n)$:

$$\mathbf{P}(C_n) = \mathbf{P}(E_{X_1}E_{X_2}\cdots E_{X_n}) = \mathbf{P}(E_{X_1})\mathbf{P}(E_{X_2}|E_{X_1})\cdots\mathbf{P}(E_{X_n}|E_{X_{n-1}})$$
$$= \pi_{X_1}p_{X_1X_2}\cdots p_{X_{n-1}X_n} = \pi_{X_1}\prod_{(k,l)} p_{kl}^{m_{(k,l)}(n-1)}.$$

The product here is taken over all pairs (k,l). Therefore ($\pi_i = \mathbf{P}(X_1 = i)$)

$$\log\mathbf{P}(C_n) = \log\pi_{X_1} + \sum_{k,l} m_{(k,l)}(n-1)\log p_{kl},$$
$$\frac{1}{n}\log\mathbf{P}(C_n) \xrightarrow{p} \sum_{k,l}\pi_k p_{kl}\log p_{kl} = -H. \qquad \square$$

14.2.3 The Asymptotic Behaviour of the Number of the Most Common Outcomes in a Sequence of Trials

Theorem 14.2.1 has an important corollary. Rank all the messages (words) C_n of length n according to the values of their probabilities in descending order. Next pick the most probable words one by one until the sum of their probabilities exceeds a prescribed level α, $0 < \alpha < 1$. Denote the number (and also the set) of the selected words by $M_\alpha(n)$.

Theorem 14.2.2 *For each* $0 < \alpha < 1$, *there exists one and the same limit*

$$\lim_{n\to\infty}\frac{\log M_\alpha(n)}{n} = H.$$

Proof Let $\delta > 0$ be a number, which can be arbitrarily small. We will say that C_n falls into category K_1 if its probability $\mathbf{P}(C_n) > 2^{-n(H-\delta)}$, and into category K_2 if

$$2^{-n(H+\delta)} < \mathbf{P}(C_n) \le 2^{-n(H-\delta)}.$$

Finally, C_n belongs to the third category K_3 if

$$\mathbf{P}(C_n) \le 2^{-n(H+\delta)}.$$

Since, by Theorem 14.2.1, $\mathbf{P}(C_n \in K_1 \cup K_3) \to 0$ as $n \to \infty$, the set $M_\alpha(n)$ contains only the words from K_1 and K_2, and the last word from $M_\alpha(n)$ (i.e. having the smallest probability)—we denote it by $C_{\alpha,n}$—belongs to K_2. This means that

$$M_\alpha(n)2^{-n(H+\delta)} < \sum_{C_n\in M_\alpha(n)}\mathbf{P}(C_n) < \alpha + \mathbf{P}(C_{\alpha,n}) < \alpha + 2^{-n(H-\delta)}.$$

This implies

$$\frac{\log M_\alpha(n)}{n} < \frac{(\alpha + 2^{-n(H-\delta)})}{n} + H + \delta.$$

Since δ is arbitrary, we have

$$\limsup_{n\to\infty} \frac{\log M_\alpha(n)}{n} \le H.$$

On the other hand, the words from K_2 belonging to $M_\alpha(n)$ have total probability $\ge \alpha - \mathbf{P}(K_1)$. If $M_\alpha^{(2)}(n)$ is the number of these messages then

$$M_\alpha^{(2)}(n)2^{-n(H-\delta)} \ge \alpha - \mathbf{P}(K_1),$$

and, consequently,

$$M_\alpha(n)2^{-n(H-\delta)} \ge \alpha - \mathbf{P}(K_1).$$

Since $\mathbf{P}(K_1) \to 0$ as $n \to \infty$, for sufficiently large n one has

$$\frac{\log M_\alpha(n)}{n} \ge H - \delta + \frac{1}{n}\log\frac{\alpha}{2}.$$

It follows that

$$\limsup_{n\to\infty} \frac{\log M_\alpha(n)}{n} \ge H.$$

The theorem is proved. □

Now one can obtain a useful interpretation of this theorem. Let N be the number of the chain states. Suppose for simplicity's sake that $N = 2^m$. Then the number of different words of length n (chains C_n) will be equal to $N^n = 2^{nm}$. Suppose, further, that these words are transmitted using a binary code, so that m binary symbols are used to code every state. Thus, with such transmission method—we will call it *direct coding*—the length of the messages will be equal to nm. (For example, one can use Markov chains to model the Russian language and take $N = 32$, $m = 5$.) The assertion of Theorem 14.2.2 means that, for large n, with probability $1 - \varepsilon$, $\varepsilon > 0$, only 2^{nH} of the totality of 2^{nm} words will be transmitted. The probability of transmitting all the remaining words will be small if ε is small. From this it is easy to establish the existence of another more economical code requiring, with a large probability, a smaller number of digits to transmit a word. Indeed, one can enumerate the selected 2^{nH} most likely words using, say, a binary code again, and then transmit only the number of the word. This clearly requires only nH digits. Since we always have $H \le \log N = m$, the length of the message will be $m/H \ge 1$ times smaller.

This is a special case of the so-called *basic coding theorem* for Markov chains: for large n, there exists a code for which, with a high probability, the original message C_n can be transmitted by a sequence of signals which is m/H times shorter than in the case of the direct coding.

The above coding method is rather an oversimplified example than a recipe for efficiently compressing the messages. It should be noted that finding a really efficient coding method is a rather difficult task. For example, in Morse code it is reasonable to encode more frequent letters by shorter sequences of dots and dashes.

However, the text reduction by m/H times would not be achieved. Certain compression techniques have been used in this book as well. For example, we replaced the frequently encountered words "characteristic function" by "ch.f." We could achieve better results if, say, shorthand was used. The structure of a code with a high compression coefficient will certainly be very complicated. The theorems of the present chapter give an upper bound for the results we can achieve.

Since $H = \sum \frac{1}{n} \log N = m$, for a sequence of *independent* equiprobable symbols, such a text is *incontractible*. This is why the proximity of "new" messages (encoded using a new alphabet) to a sequence of equiprobable symbols could serve as a criterion for constructing new codes.

It should be taken into account, however, that the text "redundancy" we are "fighting" with is in many cases a useful and helpful phenomenon. Without such redundancy, it would be impossible to detect misprints or reconstruct omissions as easily as we, say, restore the letter "r" in the word "info · mation".

The reader might know how difficult it is to read a highly abridged and formalised mathematical text. While working with an ideal code no errors would be admissible (even if we could find any), since it is impossible to reconstruct an omitted or distorted symbol in a sequence of equiprobable digits. In this connection, there arises one of the basic problems of information theory: to find a code with the smallest "redundancy" which still allows one to eliminate the transmission noise.