Chapter 11
Entropy and Information Theory

Von Neumann told me, "You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage."

C.E. Shannon (1916–2001)

11.1 Introduction

The term entropy was first used by R. Clausius in 1865, in the setting of his research on heat. The underlying concept would play a crucial role in the development of thermodynamics and statistical mechanics with the work of J.W. Gibbs and L. Boltzmann at the end of the nineteenth century. It was, however, not these two theories that inspired A.N. Kolmogorov when he introduced a new invariant called "entropy" to study dynamical systems, but rather the work of C.E. Shannon on information theory.

In a well-known paper published in 1948, marking the birth of information theory, C.E. Shannon introduced a quantity meant to quantify the information lost in telephone transmissions when there is static on the line.

The following experiment shows how we can understand entropy from the point of view of information theory. Consider an information source that produces a value or result belonging to a set of n symbols $x_1, \ldots, x_n$ with respective probabilities $p_1, \ldots, p_n$. We wish to determine the result produced by the information source by asking yes-no questions, of the type "Is the result equal to x_1?" or "Does the result belong to such and such a subset?". Set $H = -\sum p_i \log p_i$, where the logarithm is in base 2. C.E. Shannon showed that the average number of questions necessary lies between H and $H + 1$ provided that the choice of the questions is optimal.

In Chap. 10, we introduced the concept of partition and associated information function. What is the connection with the question we just stated? First, note that

Y. Coudène, *Ergodic Theory and Dynamical Systems*, Universitext,
DOI 10.1007/978-1-4471-7287-1_11

a question partitions the sample space into two subsets. A sequence of questions therefore gives a sequence of partitions of our space. If this sequence of questions is able to distinguish between all possible results, this means that the generated partition is the partition into singletons, where $\xi(x) = \{x\}$ for all x. The entropy of this partition is exactly equal to H.

The value of H manifests as the average amount of information needed to determine the result produced by the source. Later on, we will explain how to quantify this concept of information and treat a number of concrete examples. We will compute explicitly the entropy H when the source produces a sequence of mutually independent symbols or when the probability of a symbol depends only on the symbol that precedes it (Markov case).

11.2 The Notion of Information

Alice is informed of the result of a random experiment. Bob wants to determine this result and asks Alice to give him information. Alice allows Bob to ask one question, which she will answer with "yes" or "no".

Bob asks his question, and Alice gives a positive answer. How much information has Bob received?

Let us try to understand the value of this information. We first note that Bob's question partitions the sample space Ω into two subsets: on the one hand, the results that lead to a positive answer from Alice and on the other hand, those that lead to a negative answer.

The amount of information received by Bob depends on the probability p of obtaining a positive answer to his question; we denote this amount by $I(p)$. Next, suppose that Alice carries out the experiment twice and that the answer to the first question is positive in both cases, and let us state the following postulate.

The value of the information provided by the two results obtained in an independent manner is equal to the sum of the amount of information associated with each of the results.

The amount of information obtained by Bob is therefore equal to twice the amount that would have resulted from a positive answer to a single execution of the experiment. On the other hand, the probability of obtaining a positive answer twice is equal to p^2. So we have $I(p^2) = 2I(p)$. It follows that $I(p^m) = mI(p)$ when we repeat the experiment m times. If the function $p \mapsto I(p)$ is continuous on $(0, 1)$, this leads to the equality $I(p^x) = xI(p)$ for every real number x. By convention, we set $I(\frac{1}{2}) = 1$, which gives $I(y) = -\log_2(y)$.

Let us compute the average amount of information given by Alice's answer: the answer is positive with probability p, in which case the amount of information

received is equal to $\log_2 p$; the answer is negative with probability $1-p$, in which case the amount of information received is equal to $\log_2(1-p)$. We therefore have

$$H(p) = -p\log_2 p - (1-p)\log_2(1-p).$$

Denote by $\xi = \{yes, no\}$ the partition of Ω associated with Alice's question, by x the result of the random experiment, and by μ the probability measure defined on Ω. By the definition of p, we have $p = P(yes) = \mu(\xi(x))$, and we recover the usual formula for the entropy of a partition with two elements. Let us now treat a concrete example Sect. 11.3.

11.3 The Game of Questions and Answers

Alice rolls two six-sided dice and takes the sum of the outcomes (Table 11.1). Bob tries to guess the result by asking questions to which Alice replies with "yes" or "no".

Bob asks, for example, whether the result is greater than or equal to 7, then whether it is even. The respective answers to these questions are "no" and "yes". He then asks whether the result is equal to 6, and after receiving a negative answer, whether it is equal to 2. This is summarized in Table 11.2. We can calculate explicitly the information given by Alice's replies. Denote Alice's result by x, and by $\xi_1,\ldots,\xi_n$ the partitions associated with Bob's successive questions. The information Bob obtains from the answers to questions 1 through n is equal to $I(\xi_1 \vee \xi_2 \vee \cdots \vee \xi_n)(x)$; it can be found in the penultimate column of the table.

The information gain given by answer n is $I(\xi_n \mid \xi_1 \vee \cdots \vee \xi_{n-1})(x)$, that is, $-\log_2 \mu(\xi_n(x) \mid \xi_1 \vee \cdots \vee \xi_{n-1}(x))$. This is the difference between $I(\xi_1 \vee \xi_2 \vee \cdots \vee \xi_n)(x)$ and $I(\xi_1 \vee \xi_2 \vee \cdots \vee \xi_{n-1})(x)$; it can be found in the last column of the table.

To guess the result x, Bob must obtain a total gain of information equal to $-\log_2\big(P(\{x\})\big)$. We can follow his progress in the table. For example, the reply to question 3, "*Is the result equal to* 6*?*", is rather favorable (information greater than 1), even if on average such a question brings little information in this context (the average relative information is equal to 0.25). After having asked question 4, "*Is the result equal to* 2*?*", Bob has enough information to guess the number. Indeed, 4 is the only result that induces the series of answers "no-yes-no-no" to his questions, and in fact, Bob has reached the necessary amount of information: $\log_2(12) \simeq 3.58$.

11.4 Information and Markov Chains

Here is another example from probability theory. Let $X_0, X_1, \ldots, X_n, \ldots$ be a sequence of stationary random variables on a probability space $(\Omega, \mathcal{T}, P)$, with

Table 11.1 Probability distribution of the sum of the outcomes of rolling two six-sided dice

Result x	2	3	4	5	6	7	8	9	10	11	12
Probability p	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
Information	5.17	4.17	3.58	3.17	2.85	2.58	2.85	3.17	3.58	4.17	5.17

Table 11.2 List of questions and answers ($x = 4$)

i	Question	An element of ξ_i	Cumulative entropy	Gain	Answer	Cumulative information	Gain
1	$\geqslant 7$?	$\{7, 8, 9, 10, 11, 12\}$	0.98	0.98	No	1.26	1.26
2	Even?	$\{2, 4, 6, 8, 10, 12\}$	1.96	0.98	Yes	2	0.74
3	6?	$\{6\}$	2.21	0.25	No	3.17	1.17
4	2?	$\{2\}$	2.3	0.09	No	3.58	0.41

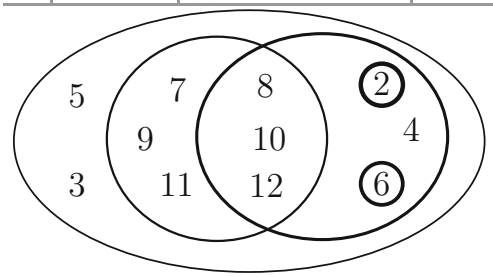
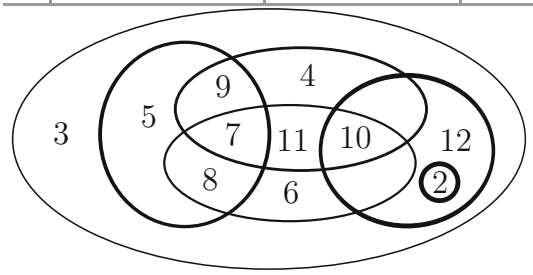


Table 11.3 List of questions and answers ($x = 12$)

i	Question	An element of ξ_i	Cumulative entropy	Gain	Answer	Cumulative information	Gain
1	6, 7, 8, 10, or 11?	$\{6, 7, 8, 10, 11\}$	0.98	0.98	No	1.26	1.26
2	4, 7, 9, 10, or 11?	$\{4, 7, 9, 10, 11\}$	1.98	1	No	2.16	0.9
3	5, 7, 8, or 9?	$\{5, 7, 8, 9\}$	2.97	0.99	No	3.18	1.02
4	2, 10, or 12?	$\{2, 10, 12\}$	3.22	0.25	Yes	4.17	0.99
5	2?	$\{2\}$	3.28	0.06	No	5.17	1



values in the finite set $\mathcal{A} = \{1, \dots, N\}$. Suppose that we know the values of the X_i for $i \geqslant 1$. What amount of information do we need, on average, to know the value of X_0?

We may assume that the space Ω on which the X_i are defined is equal to $\mathcal{A}^{\mathbf{N}}$, where the random variable X_i corresponds to the projection onto the coordinate i. We denote by $\sigma : \Omega \to \Omega$ the shift to the left. Let us consider the partitions ξ_n of Ω defined by $\xi_n = \{(X_n = a) \mid a \in \mathcal{A}\}$; note that $\xi_n = \sigma^{-n}\xi_0$. Let $H(X_0 \mid X_1, \dots, X_n)$ be the average amount of information needed to know the value of X_0 if we know that of the X_i for $i = 1, \dots, n$. We have the equality

$$H(X_0 \mid X_1, \dots, X_n) = H(\xi_0 \mid \xi_1 \vee \xi_2 \vee \cdots \vee \xi_n) = H\big(\xi_0 \mid \bigvee_{i=1}^{n} \sigma^{-i}\xi_0\big).$$

This amount converges to the entropy of the shift $h(\sigma, \xi_0)$. When $\Omega = \mathcal{A}^{\mathbf{N}}$, the partition ξ_0 is generating because the elements of the partition $\bigvee_{0}^{n} \sigma^{-i}\xi_0$ are the cylinder sets of length $n + 1$. The entropy of the shift can therefore be seen as the average amount of information needed to know the "current" value X_0 if we know the "past" values X_i of the process for $i \geqslant 1$. Let us carry out the calculation when the X_i form a Markov chain.

Proposition 11.1 *Let $\mathcal{A}$ be a finite set; the shift $\sigma : \mathcal{A}^{\mathbf{N}} \to \mathcal{A}^{\mathbf{N}}$ is defined by the formula $\sigma(\{x_i\}_{i\in\mathbf{N}}) = \{x_{i+1}\}_{i\in\mathbf{N}}$. For $i, j \in \mathcal{A}$, we consider elements p_i, $p_{i,j}$ of $[0, 1]$ satisfying $\sum_i p_i = 1$, $\sum_j p_{i,j} = 1$, and $\sum_i p_i p_{i,j} = p_j$. Denote by P the probability satisfying*

$$P\big(\{\{x_i\}_{i\in\mathbf{N}} \mid x_0 = i_0, \dots, x_n = i_n\}\big) = p_{i_0} p_{i_0,i_1} p_{i_1,i_2} \cdots p_{i_{n-1},i_n}.$$

The entropy of σ with respect to P is given by

$$h_P(\sigma) = -\sum_{i,j} p_i p_{i,j} \log_2 p_{i,j}.$$

Proof By virtue of the equalities $P(X_k = j \mid X_{k+1} = i) = p_{i,j}$ and $p_i = P(X_1 = i)$ and the Markov property, we have

$$\begin{aligned} H(X_0 \mid X_1, \dots, X_n) &= H(X_0 \mid X_1) \\ &= -\sum_{i,j} P(X_0 = j, X_1 = i) \log_2 P(X_0 = j \mid X_1 = i) \\ &= -\sum_{i,j} p_i p_{i,j} \log_2 p_{i,j}. \end{aligned}$$

□

As a corollary, we deduce that a Bernoulli shift on an alphabet with n symbols with respective probabilities $p_1, \dots, p_n$ has entropy $-\sum p_i \log_2 p_i$. For example, the Bernoulli shift corresponding to flipping a fair coin ($p_1 = p_2 = \frac{1}{2}$) has entropy equal to $\log_2 2 = 1$. The Bernoulli shift corresponding to rolling a six-sided die ($p_1 = \cdots = p_6 = \frac{1}{6}$) has entropy $\log_2 6$.

In Chap. 10, we saw that two measure-preserving dynamical systems with different entropies cannot be isomorphic.

Corollary 11.1 *Two Bernoulli shifts with different entropies are not isomorphic. In particular, the Bernoulli shift with probability vector* $(\frac{1}{2}, \frac{1}{2})$ *is not isomorphic to the Bernoulli shift with probability vector* $(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$.

11.5 Interpretation in the Dynamical Setting

Consider a transformation T that admits a one-sided generating partition ξ, in the sense that the elements of the $T^{-i}\xi$ for $i \in \mathbf{N}$ generate the σ-algebra of measurable sets. Let us try to interpret the entropy $h(T)$ of T in terms of information. We saw in Chap. 10 that this entropy is given by the formula $h(T) = h(T, \xi) = H(\xi \mid T^{-1}\mathcal{T})$; it corresponds to the average amount of information needed to determine to which element of ξ the point x belongs if we know the positions of the iterates $T^i(x)$ in the partition ξ for $i \geqslant 1$.

Under very general hypotheses, we can show that ξ is generating if and only if the sequence $\xi(T^i(x))$ for $i \geqslant 0$ determines x uniquely if x belongs to a certain well-chosen set of full measure. To formalize this result, we need preliminaries on measure theory that are the object of the last part of this book.

Proposition 11.2 *Let* $(X, \mathcal{T}, \mu)$ *be a Lebesgue space, and let* $T : X \to X$ *be a measurable map that preserves the measure* μ. *A finite partition* ξ *is one-sided generating if and only if there exists a set* $X_0 \subset X$ *of full measure such that the map*

$$\begin{aligned} X &\longrightarrow \xi^{\mathbf{N}} \\ x &\longmapsto \{\xi(T^i(x))\}_{i \in \mathbf{N}} \end{aligned}$$

restricted to X_0 *is injective.*

In other words, once the points of X_0^c have been dismissed, the position of the point x is fixed if we know the list of elements of ξ to which x and its iterates belong. This result will be proved in Chap. 15. We use the notion of Lebesgue space, a notion that will be explained in Part IV of this book. All probability spaces one comes across in practice are Lebesgue spaces. For example, every Borel space of a complete metric space, endowed with a Borel probability measure, is a Lebesgue space.

Using this result, we can interpret the entropy of a transformation in terms of information. Knowing the position of all the iterates $T^i(x)$ for $i > 0$ with respect to the generating partition ξ corresponds to knowing the point $T(x)$, and the only thing missing to determine the point x completely is the position of x with respect to the partition itself. The entropy $h(T)$ can therefore be seen as the average amount of information needed to know x if we know $T(x)$.

If the transformation T is invertible, the point x is completely determined by the knowledge of $T(x)$, and the amount of information needed to know x if we know $T(x)$ is 0. The entropy of an invertible transformation with a one-sided generating partition is 0. This does not mean that all invertible transformations have entropy 0, rather that in general, these transformations do not have one-sided generating partitions, whence the need to turn to generating partitions to calculate their entropy.

11.6 Exercises

11.6.1 Basic Exercises

***Exercise* 1** Alice rolls two six-sided dice and takes the sum of the outcomes. She agrees to answer Bob's questions about the value of the sum with "yes" or "no".

Can Bob be certain to guess the correct value using only three questions? What is the minimal number of questions he must ask to be certain to conclude regardless of the result? Repeat this exercise for three and then four dice.

***Exercise* 2** Alice rolls two six-sided dice and Bob tries to guess the sum of the outcomes. He is allowed to ask four questions.

Alice's first three answers have given him an amount of information equal to 3.17, and he only has one question left.

- Can the result be 6?
- Can the result be 4?
- If so, which question should Bob ask?

Recall that $\log_2(3) = 1.58$ *and* $\log_2(5) = 2.32$.

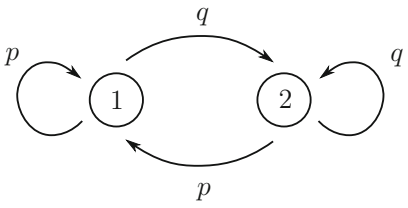
***Exercise* 3** Let $X_0, X_1, \ldots, X_n, \ldots$ be a stationary Markov sequence of random variables. Show the inequality $H(X_n \mid X_0) \leqslant H(X_{n+1} \mid X_0)$.

***Exercise* 4** We roll two n-sided dice and take the sum of the outcomes. Calculate the entropy of the system obtained by repeating this experiment independently. Compare this with the entropy associated with the independent repetition of a uniformly distributed experiment on a set with $2n - 1$ elements.

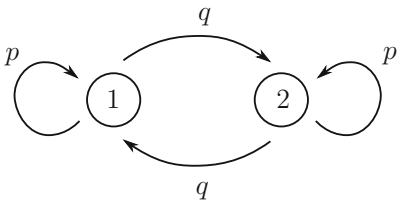
***Exercise* 5** Show that the two-sided Bernoulli shift (where "two-sided" means indexed by $\mathbf{Z}$) on an alphabet with three symbols with respective probabilities $\frac{1}{3}$,

$\frac{1}{3}$, and $\frac{1}{3}$ is not isomorphic to the Bernoulli shift on an alphabet with two symbols with respective probabilities $\frac{1}{2}$ and $\frac{1}{2}$.

***Exercise* 6** Let $p, q \in [0, 1]$ satisfy $p+q = 1$. Compute the entropy of the following Markov chains:



Transition matrix: $\begin{pmatrix} p & q \\ p & q \end{pmatrix}$



Transition matrix: $\begin{pmatrix} p & q \\ q & p \end{pmatrix}$

11.6.2 More Advanced Exercise

***Exercise* 7** We consider a random experiment with n possible outcomes with respective probabilities $p_1, \ldots, p_n$. Bob tries to guess the outcome of the experiment using only yes-no questions. Show that the minimal number of questions Bob needs to ask to be certain to conclude regardless of the result is always greater than the entropy $-\sum p_i \log_2 p_i$. Show that equality is possible only if all outcomes of the experiment have the same probability. The number of possible outcomes must then be a power of 2.

11.7 Comments

Let us return to the case of a six-sided die. Consider a set of questions $\xi_1, \ldots, \xi_n$ that allows us to conclude regardless of the outcome: $\xi_1 \vee \cdots \vee \xi_n(x) = \{x\}$, that is, $H(\xi_1 \vee \cdots \vee \xi_n) = -\sum p_i \log p_i = 3.27$. For some outcomes, it is not necessary to ask n questions to conclude. For example, for the questions in Table 11.3, it suffices to ask the first three to find the result if it is 4, 5, 6, 7, 8, or 9.

On average, what is the number of questions truly asked to find the outcome? In his fundamental paper of 1948, C.E. Shannon showed that this average number is always greater than the entropy. In 1952, D. Huffman proposed an algorithm to construct a sequence of questions that minimizes the average number of questions that need to be asked. For the rolling of two dice, Table 11.3 was obtained using this algorithm. The average number of questions truly asked is 3.306; this is optimal. The compression methods jpeg, mp3, and pkzip use this algorithm by D. Huffman.

Note that at least four questions need to be asked to distinguish between all outcomes. Indeed, three questions partition the set of outcomes into at most $2^3 = 8$ parts, whereas there are 11 different outcomes.

We can try to determine a set of four questions such that the average number of questions that need to be asked out of this set of four is minimal. This can be obtained using a "numismatic" algorithm, which gives the questions "6, 7, 8, or 9?"; "4, 5, 7, 8, or 10?"; "3, 5, 6, 7, or 11?"; and "2, 3, or 4?". The average number of questions needed is 3.333.
The entropy associated with rolling two n-sided dice is given by

$$H = -\sum p_i \log p_i = 2\log n - \frac{1}{n^2}\Big(\sum_{i=1}^{n-1} 2i \log i + n \log n\Big) \sim \log(n) + \frac{1}{2} + o(1/n).$$

The reader may want to compare this with the entropy of the uniform distribution on a set with $2n-1$ elements: $H = \log(2n-1) \sim \log(n) + \log(2) + o(1)$.