

# 9

## Entropy

The word *entropy* was invented in 1865 by the German physicist and mathematician Rudolf Clausius, one of the founding pioneers of thermodynamics. In the theory of systems in thermodynamical equilibrium, the entropy quantifies the degree of “disorder” in the system. The second law of thermodynamics states that, when an isolated system passes from an equilibrium state to another, the entropy of the final state is necessarily bigger than the entropy of the initial state. For example, when we join two containers with different gases (oxygen and nitrogen, say), the two gases mix with one another until reaching a new macroscopic equilibrium, where they are both uniformly distributed in the two containers. The entropy of the new state is larger than the entropy of the initial equilibrium, where the two gases were separate.

The notion of entropy plays a crucial role in different fields of science. An important example, which we explore in our presentation, is the field of information theory, initiated by the work of the American electrical engineer Claude Shannon in the mid 20th century. At roughly the same time, the Russian mathematicians Andrey Kolmogorov and Yakov Sinai were proposing a definition of the entropy of a system in ergodic theory. The main purpose was to provide an invariant of ergodic equivalence that, in particular, could distinguish two Bernoulli shifts. This Kolmogorov–Sinai entropy is the subject of this chapter.

In Section 9.1 we define the entropy of a transformation with respect to an invariant probability measure, by analogy with a similar notion in information theory. The theorem of Kolmogorov–Sinai, which we discuss in Section 9.2, is a fundamental tool for the actual calculation of the entropy in specific systems. In Section 9.3 we analyze the concept of entropy from a more local viewpoint, which is more closely related to Shannon’s formulation of this concept. Next, in Section 9.4, we illustrate a few methods for calculating the entropy, by means of concrete examples.

In Section 9.5 we discuss the role of the entropy as an invariant of ergodic equivalence. The highlight is the theorem of Ornstein (Theorem 9.5.2), according to which any two-sided Bernoulli shifts are ergodically equivalent

if and only if they have the same entropy. In that section we also introduce the class of Kolmogorov systems, which contains the Bernoulli shifts and is contained in the class of systems with Lebesgue spectrum. In both cases the inclusion is strict.

In the last couple of sections we present two complementary topics that will be useful later. The first one (Section 9.6) is the theorem of Jacobs, according to which the entropy behaves in an affine way with respect to the ergodic decomposition. The other (Section 9.7) concerns the notion of the Jacobian and its relations with the entropy.

## 9.1 Definition of entropy

To motivate the definition of Kolmogorov–Sinai entropy, let us look at the following basic situation in information theory. Consider some communication channel transmitting symbols from a certain alphabet  $\mathcal{A}$ , one after the other. This could be a telegraph transmitting group of dots and dashes, according to the old Morse code, an optical fiber, transmitting packets of zeros and ones, according to the ASCII binary code, or any other process of sequential transmission of information, such as our reader’s going through the text of this book, one letter after the other. The objective is to measure the *entropy* of the channel, that is, the mean *quantity of information* it carries, per unit of time.

### 9.1.1 Entropy in information theory

It is assumed that each symbol has a given frequency, that is, a given probability of being used at any time in the communication. For example, if the channel is transmitting a text in English then the letter  $E$  is more likely to be used than the letter  $Z$ , say. The occurrence of rarer symbols, such as  $Z$ , restricts the kind of word or sentence in which they appear and, hence, is more informative than the presence of commoner symbols, such as  $E$ .

This suggests that *information* should be a function of *probability*: the more unlikely a symbol (or a *word*, defined as a finite sequence of symbols) is, the more information it carries.

The situation is actually more complicated, because for most communication codes the probability of using a given symbol also depends on the context. For example, still assuming that the channel transmits in English, any sequence of symbols  $S, Y, S, T, E$  must be followed by an  $M$ : in this case, in view of the symbols transmitted previously, this letter  $M$  is unavoidable, which also means that it carries no additional information.<sup>1</sup>

<sup>1</sup> We once participated in a “treasure hunt” that consisted in searching the woods for hidden letters that would form the name of a mathematical object. It just so happened that the first three letters that were found were  $Z, Z$  and  $Z$ . That unfortunate circumstance ended the game

On the other hand, in those situations where symbols are transmitted at random, independently of each other, the information carried by each symbol simply adds to the information conveyed by the previous ones. For example, if the transmission reflects the outcomes of the successive flipping of a fair coin, then the amount of information associated with the outcome (*Head, Tail, Tail*) must be equal to the *sum* of the amounts of information associated with each of the symbols *Head, Tail* and *Tail*. Now, by independence, the probability of the event (*Head, Tail, Tail*) is the *product* of the probabilities of the events *Head, Tail* and *Tail*.

This suggests that information should be defined in terms of the *logarithm* of the probability. In information theory it is usual to consider base 2 logarithms, because essentially all the communication channels one finds in practice are binary. However, there is no reason to stick to that custom in our setting: we will consider natural (base *e*) logarithms instead.

By definition, the *quantity of information* associated with a symbol  $a \in \mathcal{A}$  is given by

$$I(a) = -\log p_a, \tag{9.1.1}$$

where  $p_a$  is the probability (frequency) of the symbol  $a$ . The *mean information* associated with the alphabet  $\mathcal{A}$  is given by

$$I(\mathcal{A}) = \sum_{a \in \mathcal{A}} p_a I(a) = \sum_{a \in \mathcal{A}} -p_a \log p_a. \tag{9.1.2}$$

More generally, the quantity of information associated with a word  $a_1 \dots a_n$  is

$$I(a_1 \dots a_n) = -\log p_{a_1 \dots a_n}, \tag{9.1.3}$$

where  $p_{a_1 \dots a_n}$  denotes the probability of the word. In the independent case this coincides with the product  $p_{a_1} \dots p_{a_n}$  of the probabilities of the symbols, but not in general. Denoting by  $\mathcal{A}^n$  the set of all the words of length  $n$ , we define

$$I(\mathcal{A}^n) = \sum_{a_1, \dots, a_n} p_{a_1 \dots a_n} I(a_1, \dots, a_n) = \sum_{a_1, \dots, a_n} -p_{a_1 \dots a_n} \log p_{a_1 \dots a_n}. \tag{9.1.4}$$

Finally, the *entropy* of the communication channel is defined by:

$$I = \lim_n \frac{1}{n} I(\mathcal{A}^n). \tag{9.1.5}$$

We invite the reader to check that the sequence  $I(\mathcal{A}^n)$  is subadditive and, thus, the limit in (9.1.5) does exist. This is also contained in the much more general theory that we are about to present.

prematurely, since at that point the remaining letters could add no information: there is only one mathematical object whose name includes the letter Z three times (the *Yoccoz puzzle*).

### 9.1.2 Entropy of a partition

We want to adapt these ideas to our context in ergodic theory. The main difference is that, while in information theory the alphabet  $\mathcal{A}$  is usually discrete (finite or, at most, countable), that is not the case for the domain (space of states) of most interesting dynamical systems. That issue is dealt with by using partitions of the domain.

Let  $(M, \mathcal{B}, \mu)$  be a probability space. In this chapter, by *partition* we always mean a **countable** (finite or infinite) family  $\mathcal{P}$  of pairwise disjoint measurable subsets of  $M$  whose union has full measure. We denote by  $\mathcal{P}(x)$  the element of the partition that contains a given point  $x$ . The *sum*  $\mathcal{P} \vee \mathcal{Q}$  of two partitions  $\mathcal{P}$  and  $\mathcal{Q}$  is the partition whose elements are the intersections  $P \cap Q$  with  $P \in \mathcal{P}$  and  $Q \in \mathcal{Q}$ . More generally, given any countable family of partitions  $\mathcal{P}_n$ , we define

$$\bigvee_n \mathcal{P}_n = \left\{ \bigcap_n P_n : P_n \in \mathcal{P}_n \text{ for each } n \right\}.$$

With each partition  $\mathcal{P}$  we associate the corresponding *information function*

$$I_{\mathcal{P}} : M \rightarrow \mathbb{R}, \quad I_{\mathcal{P}}(x) = -\log \mu(\mathcal{P}(x)). \quad (9.1.6)$$

It is clear that the function  $I_{\mathcal{P}}$  is measurable. By definition, the *entropy* of the partition  $\mathcal{P}$  is the mean of its information function, that is,

$$H_{\mu}(\mathcal{P}) = \int I_{\mathcal{P}} d\mu = \sum_{P \in \mathcal{P}} -\mu(P) \log \mu(P). \quad (9.1.7)$$

We always abide to the usual (in the theory of Lebesgue integration) convention that  $0 \log 0 = \lim_{x \rightarrow 0} x \log x = 0$ . See Figure 9.1.

Consider the function  $\phi : (0, \infty) \rightarrow \mathbb{R}$  given by  $\phi(x) = -x \log x$ . One can readily check that  $\phi'' < 0$ . Therefore,  $\phi$  is strictly concave:

$$t_1 \phi(x_1) + \cdots + t_k \phi(x_k) \leq \phi(t_1 x_1 + \cdots + t_k x_k) \quad (9.1.8)$$

for every  $x_1, \dots, x_k > 0$  and  $t_1, \dots, t_k > 0$  with  $t_1 + \cdots + t_k = 1$ ; moreover, the identity holds if and only if  $x_1 = \cdots = x_k$ . This observation will be useful on several occasions.

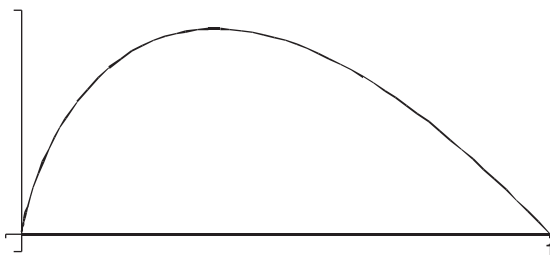


Figure 9.1. Graph of the function  $\phi(x) = -x \log x$

We say that two partitions  $\mathcal{P}$  and  $\mathcal{Q}$  are *independent* if  $\mu(P \cap Q) = \mu(P)\mu(Q)$  for every  $P \in \mathcal{P}$  and every  $Q \in \mathcal{Q}$ . Then,  $I_{\mathcal{P} \vee \mathcal{Q}} = I_{\mathcal{P}} + I_{\mathcal{Q}}$  and, therefore,  $H_{\mu}(\mathcal{P} \vee \mathcal{Q}) = H_{\mu}(\mathcal{P}) + H_{\mu}(\mathcal{Q})$ . In general, one has the inequality  $\leq$ , as we are going to see in a while.

**Example 9.1.1.** Let  $M = [0, 1]$  be endowed with the Lebesgue measure. For each  $n \geq 1$ , consider the partition  $\mathcal{P}^n$  of the interval  $M$  into the subintervals  $((i - 1)/10^n, i/10^n]$  with  $1 \leq i \leq 10^n$ . Then,

$$H_{\mu}(\mathcal{P}^n) = \sum_{i=1}^{10^n} -10^{-n} \log 10^{-n} = n \log 10.$$

**Example 9.1.2.** Let  $M = \{1, \dots, d\}^{\mathbb{N}}$  be endowed with a product measure  $\mu = \nu^{\mathbb{N}}$ . Denote  $p_i = \nu(\{i\})$  for each  $i \in \{1, \dots, d\}$ . For each  $n \geq 1$ , let  $\mathcal{P}^n$  be the partition of  $M$  into the cylinders  $[0; a_1, \dots, a_n]$  of length  $n$ . The entropy of  $\mathcal{P}^n$  is

$$\begin{aligned} H_{\mu}(\mathcal{P}^n) &= \sum_{a_1, \dots, a_n} -p_{a_1} \dots p_{a_n} \log(p_{a_1} \dots p_{a_n}) \\ &= \sum_j \sum_{a_1, \dots, a_n} -p_{a_1} \dots p_{a_j} \dots p_{a_n} \log p_{a_j} \\ &= \sum_j \sum_{a_j} -p_{a_j} \log p_{a_j} \sum_{a_i, i \neq j} p_{a_1} \dots p_{a_{j-1}} p_{a_{j+1}} \dots p_{a_n}. \end{aligned}$$

The last sum is equal to 1, since  $\sum_i p_i = 1$ . Therefore,

$$H_{\mu}(\mathcal{P}^n) = \sum_{j=1}^n \sum_{a_j=1}^d -p_{a_j} \log p_{a_j} = \sum_{j=1}^n \sum_{i=1}^d -p_i \log p_i = -n \sum_{i=1}^d p_i \log p_i.$$

**Lemma 9.1.3.** *Every finite partition  $\mathcal{P}$  has finite entropy:  $H_{\mu}(\mathcal{P}) \leq \log \#\mathcal{P}$  and the identity holds if and only if  $\mu(P) = 1/\#\mathcal{P}$  for every  $P \in \mathcal{P}$ .*

*Proof.* Let  $\mathcal{P} = \{P_1, P_2, \dots, P_n\}$  and consider  $t_i = 1/n$  and  $x_i = \mu(P_i)$ . By the concavity property (9.1.8):

$$\frac{1}{n} H_{\mu}(\mathcal{P}) = \sum_{i=1}^n t_i \phi(x_i) \leq \phi\left(\sum_{i=1}^n t_i x_i\right) = \phi\left(\frac{1}{n}\right) = \frac{\log n}{n}.$$

Therefore,  $H_{\mu}(\mathcal{P}) \leq \log n$ . Moreover, the identity holds if and only if  $\mu(P_i) = 1/n$  for every  $i = 1, \dots, n$ .

**Example 9.1.4.** Let  $M = [0, 1]$  be endowed with the Lebesgue measure  $\mu$ . Observe that the series  $\sum_{k=1}^{\infty} 1/(k(\log k)^2)$  is convergent. Let  $c$  be the value of the sum. Then, we may partition  $[0, 1]$  into intervals  $P_k$  with  $\mu(P_k) = 1/(ck(\log k)^2)$  for every  $k$ . Let  $\mathcal{P}$  be the partition formed by these subintervals. Then,

$$H_{\mu}(\mathcal{P}) = \sum_{k=1}^{\infty} \frac{\log c + \log k + 2 \log \log k}{ck(\log k)^2}.$$

By the ratio convergence criterion, the series on the right-hand side has the same behavior as the series  $\sum_{k=1}^{\infty} 1/(k \log k)$  which, as we know (use the integral convergence criterion), is divergent. Therefore,  $H_{\mu}(\mathcal{P}) = \infty$ .

This shows that infinite partitions may have infinite entropy. From now, for the rest of the chapter, we always consider (countable) partitions with finite entropy.

The *conditional entropy* of a partition  $\mathcal{P}$  with respect to another partition  $\mathcal{Q}$  is the number

$$H_{\mu}(\mathcal{P}/\mathcal{Q}) = \sum_{P \in \mathcal{P}} \sum_{Q \in \mathcal{Q}} -\mu(P \cap Q) \log \frac{\mu(P \cap Q)}{\mu(Q)}. \tag{9.1.9}$$

Intuitively, it measures the amount of information provided by the partition  $\mathcal{P}$  in addition to the information provided by the partition  $\mathcal{Q}$ . It is clear that  $H_{\mu}(\mathcal{P}/\mathcal{M}) = H_{\mu}(\mathcal{P})$  for every  $\mathcal{P}$ , where  $\mathcal{M}$  denotes the trivial partition  $\mathcal{M} = \{M\}$ . Moreover, if  $\mathcal{P}$  and  $\mathcal{Q}$  are independent then  $H_{\mu}(\mathcal{P}/\mathcal{Q}) = H_{\mu}(\mathcal{P})$ . In general, one has the inequality  $\leq$ , as we are going to see later.

Given two partitions  $\mathcal{P}$  and  $\mathcal{Q}$ , we say that  $\mathcal{P}$  is *coarser* than  $\mathcal{Q}$  (or, equivalently,  $\mathcal{Q}$  is *finer* than  $\mathcal{P}$ ) and we write  $\mathcal{P} < \mathcal{Q}$ , if every element of  $\mathcal{Q}$  is contained in some element of  $\mathcal{P}$ , up to measure zero. The sum  $\mathcal{P} \vee \mathcal{Q}$  may also be defined as the coarsest of all the partitions  $\mathcal{R}$  such that  $\mathcal{P} < \mathcal{R}$  and  $\mathcal{Q} < \mathcal{R}$ .

**Lemma 9.1.5.** *Let  $\mathcal{P}$ ,  $\mathcal{Q}$  and  $\mathcal{R}$  be partitions with finite entropy. Then,*

- (i)  $H_{\mu}(\mathcal{P} \vee \mathcal{Q}/\mathcal{R}) = H_{\mu}(\mathcal{P}/\mathcal{R}) + H_{\mu}(\mathcal{Q}/\mathcal{P} \vee \mathcal{R})$ ;
- (ii) if  $\mathcal{P} < \mathcal{Q}$  then  $H_{\mu}(\mathcal{P}/\mathcal{R}) \leq H_{\mu}(\mathcal{Q}/\mathcal{R})$  and  $H_{\mu}(\mathcal{R}/\mathcal{P}) \geq H_{\mu}(\mathcal{R}/\mathcal{Q})$ ;
- (iii)  $\mathcal{P} < \mathcal{Q}$  if and only if  $H_{\mu}(\mathcal{P}/\mathcal{Q}) = 0$ .

*Proof.* By definition,

$$\begin{aligned} H_{\mu}(\mathcal{P} \vee \mathcal{Q}/\mathcal{R}) &= \sum_{P, Q, R} -\mu(P \cap Q \cap R) \log \frac{\mu(P \cap Q \cap R)}{\mu(R)} \\ &= \sum_{P, Q, R} -\mu(P \cap Q \cap R) \log \frac{\mu(P \cap Q \cap R)}{\mu(P \cap R)} \\ &\quad + \sum_{P, Q, R} -\mu(P \cap Q \cap R) \log \frac{\mu(P \cap R)}{\mu(R)}. \end{aligned}$$

The sum on the right-hand side may be rewritten as

$$\begin{aligned} &\sum_{S \in \mathcal{P} \vee \mathcal{R}, Q \in \mathcal{Q}} -\mu(S \cap Q) \log \frac{\mu(S \cap Q)}{\mu(S)} + \sum_{P \in \mathcal{P}, R \in \mathcal{R}} -\mu(P \cap R) \log \frac{\mu(P \cap R)}{\mu(R)} \\ &= H_{\mu}(\mathcal{Q}/\mathcal{P} \vee \mathcal{R}) + H_{\mu}(\mathcal{P}/\mathcal{R}). \end{aligned}$$

This proves part (i). Next, observe that if  $\mathcal{P} \prec \mathcal{Q}$  then

$$\begin{aligned} H_\mu(\mathcal{P}/\mathcal{R}) &= \sum_P \sum_R \sum_{Q \subset P} -\mu(Q \cap R) \log \frac{\mu(P \cap R)}{\mu(R)} \\ &\leq \sum_P \sum_R \sum_{Q \subset P} -\mu(Q \cap R) \log \frac{\mu(Q \cap R)}{\mu(R)} = H_\mu(\mathcal{Q}/\mathcal{R}). \end{aligned}$$

This proves the first half of claim (ii). To prove the second half, note that for any  $P \in \mathcal{P}$  and  $R \in \mathcal{R}$ ,

$$\frac{\mu(R \cap P)}{\mu(P)} = \sum_{Q \subset P} \frac{\mu(Q)}{\mu(P)} \frac{\mu(R \cap Q)}{\mu(Q)}.$$

It is clear that  $\sum_{Q \subset P} \mu(Q)/\mu(P) = 1$ . Therefore, by (9.1.8),

$$\phi\left(\frac{\mu(R \cap P)}{\mu(P)}\right) \geq \sum_{Q \subset P} \frac{\mu(Q)}{\mu(P)} \phi\left(\frac{\mu(R \cap Q)}{\mu(Q)}\right)$$

for every  $P \in \mathcal{P}$  and  $R \in \mathcal{R}$ . Consequently,

$$\begin{aligned} H_\mu(\mathcal{R}/\mathcal{P}) &= \sum_{P,R} \mu(P) \phi\left(\frac{\mu(R \cap P)}{\mu(P)}\right) \geq \sum_{P,R} \mu(P) \sum_{Q \subset P} \frac{\mu(Q)}{\mu(P)} \phi\left(\frac{\mu(R \cap Q)}{\mu(Q)}\right) \\ &= \sum_{Q,R} \mu(Q) \phi\left(\frac{\mu(R \cap Q)}{\mu(Q)}\right) = H_\mu(\mathcal{R}/\mathcal{Q}). \end{aligned}$$

Finally, it follows from the definition in (9.1.9) that  $H_\mu(\mathcal{P}/\mathcal{Q}) = 0$  if and only if

$$\mu(P \cap Q) = 0 \quad \text{or else} \quad \frac{\mu(P \cap Q)}{\mu(Q)} = 1$$

for every  $P \in \mathcal{P}$  and every  $Q \in \mathcal{Q}$ . In other words, either  $Q$  is disjoint from  $P$  (up to measure zero) or else  $Q$  is contained in  $P$  (up to measure zero). This means that  $H_\mu(\mathcal{P}/\mathcal{Q}) = 0$  if and only if  $\mathcal{P} \prec \mathcal{Q}$ .

In particular, taking  $\mathcal{Q} = \mathcal{M}$  in part (ii) of the lemma we get that

$$H_\mu(\mathcal{R}/\mathcal{P}) \leq H_\mu(\mathcal{R}) \quad \text{for any partitions } \mathcal{R} \text{ and } \mathcal{P}. \tag{9.1.10}$$

Moreover, taking  $\mathcal{R} = \mathcal{M}$  in part (i) we find that

$$H_\mu(\mathcal{P} \vee \mathcal{Q}) = H_\mu(\mathcal{P}) + H_\mu(\mathcal{Q}/\mathcal{P}) \leq H_\mu(\mathcal{P}) + H_\mu(\mathcal{Q}). \tag{9.1.11}$$

Let  $f : M \rightarrow N$  be a measurable transformation and  $\mu$  be a probability measure on  $M$ . Then,  $f_*\mu$  is a probability measure on  $N$ . Moreover, if  $\mathcal{P}$  is a partition of  $N$  then  $f^{-1}(\mathcal{P}) = \{f^{-1}(P) : P \in \mathcal{P}\}$  is a partition of  $M$ . By definition,

$$\begin{aligned} H_\mu(f^{-1}(\mathcal{P})) &= \sum_{P \in \mathcal{P}} -\mu(f^{-1}(P)) \log \mu(f^{-1}(P)) \\ &= \sum_{P \in \mathcal{P}} -f_*\mu(P) \log f_*\mu(P) = H_{f_*\mu}(\mathcal{P}). \end{aligned} \tag{9.1.12}$$

In particular, if  $M = N$  and the measure  $\mu$  is invariant under  $f$  then

$$H_\mu(f^{-1}(\mathcal{P})) = H_\mu(\mathcal{P}) \quad \text{for every partition } \mathcal{P}. \tag{9.1.13}$$

We also need the following continuity property:

**Lemma 9.1.6.** *Given  $k \geq 1$  and  $\varepsilon > 0$  there exists  $\delta > 0$  such that, for any finite partitions  $\mathcal{P} = \{P_1, \dots, P_k\}$  and  $\mathcal{Q} = \{Q_1, \dots, Q_k\}$ ,*

$$\mu(P_i \Delta Q_i) < \delta \text{ for every } i = 1, \dots, k \implies H_\mu(\mathcal{Q}/\mathcal{P}) < \varepsilon.$$

*Proof.* Fix  $\varepsilon > 0$  and  $k \geq 1$ . Since  $\phi : [0, 1] \rightarrow \mathbb{R}$ ,  $\phi(x) = -x \log x$  is a continuous function, there exists  $\rho > 0$  such that  $\phi(x) < \varepsilon/k^2$  for every  $x \in [0, \rho) \cup (1 - \rho, 1]$ . Let  $\delta = \rho/k$ . Given partitions  $\mathcal{P}$  and  $\mathcal{Q}$  as in the statement, denote by  $\mathcal{R}$  the partition whose elements are the intersections  $P_i \cap Q_j$  with  $i \neq j$  and also the set  $\bigcup_{i=1}^k P_i \cap Q_i$ . Note that  $\mu(P_i \cap Q_j) \leq \mu(P_i \Delta Q_i) < \delta$  for every  $i \neq j$  and

$$\mu\left(\bigcup_{i=1}^k P_i \cap Q_i\right) \geq \sum_{i=1}^k (\mu(P_i) - \mu(P_i \Delta Q_i)) > \sum_{i=1}^k (\mu(P_i) - \delta) = 1 - \rho.$$

Therefore,

$$H_\mu(\mathcal{R}) = \sum_{R \in \mathcal{R}} \phi(\mu(R)) < \#\mathcal{R} \frac{\varepsilon}{k^2} \leq \varepsilon.$$

It is clear from the definition that  $\mathcal{P} \vee \mathcal{Q} = \mathcal{P} \vee \mathcal{R}$ . Then, using (9.1.11) and (9.1.10),

$$\begin{aligned} H_\mu(\mathcal{Q}/\mathcal{P}) &= H_\mu(\mathcal{P} \vee \mathcal{Q}) - H_\mu(\mathcal{P}) = H_\mu(\mathcal{P} \vee \mathcal{R}) - H_\mu(\mathcal{P}) \\ &= H_\mu(\mathcal{R}/\mathcal{P}) \leq H_\mu(\mathcal{R}) < \varepsilon. \end{aligned}$$

This proves the lemma.

### 9.1.3 Entropy of a dynamical system

Let  $f : M \rightarrow M$  be a measurable transformation preserving a probability measure  $\mu$ . The notion of the entropy of the system  $(f, \mu)$  that we introduce in what follows is inspired by (9.1.5).

Given a partition  $\mathcal{P}$  of  $M$  with finite entropy, denote

$$\mathcal{P}^n = \bigvee_{i=0}^{n-1} f^{-i}(\mathcal{P}) \quad \text{for each } n \geq 1.$$

Observe that the element  $\mathcal{P}^n(x)$  that contains  $x \in M$  is given by:

$$\mathcal{P}^n(x) = \mathcal{P}(x) \cap f^{-1}(\mathcal{P}(f(x))) \cap \dots \cap f^{-n+1}(\mathcal{P}(f^{n-1}(x))).$$

It is clear that the sequence  $\mathcal{P}^n$  is non-decreasing, that is,  $\mathcal{P}^n \prec \mathcal{P}^{n+1}$  for every  $n$ . Therefore, the sequence of entropies  $H_\mu(\mathcal{P}^n)$  is also non-decreasing. Another important fact is that this sequence is subadditive:



**Lemma 9.1.7.**  $H_\mu(\mathcal{P}^{m+n}) \leq H_\mu(\mathcal{P}^m) + H_\mu(\mathcal{P}^n)$  for every  $m, n \geq 1$ .

*Proof.* By definition,  $\mathcal{P}^{m+n} = \bigvee_{i=0}^{m+n-1} f^{-i}(\mathcal{P}) = \mathcal{P}^m \vee f^{-m}(\mathcal{P}^n)$ . Therefore, using (9.1.11),

$$H_\mu(\mathcal{P}^{m+n}) \leq H_\mu(\mathcal{P}^m) + H_\mu(f^{-m}(\mathcal{P}^n)). \tag{9.1.14}$$

On the other hand, since the measure  $\mu$  is invariant under  $f$ , the property (9.1.13) implies that  $H_\mu(f^{-m}(\mathcal{P}^n)) = H_\mu(\mathcal{P}^n)$  for every  $m, n$ . Substituting this fact in (9.1.14) we get the conclusion of the lemma.

In view of Lemma 3.3.4, it follows from Lemma 9.1.7 that the limit

$$h_\mu(f, \mathcal{P}) = \lim_n \frac{1}{n} H_\mu(\mathcal{P}^n) \tag{9.1.15}$$

exists and coincides with the infinitum of the sequence on the left-hand side. We call  $h_\mu(f, \mathcal{P})$  the *entropy of  $f$  with respect to the partition  $\mathcal{P}$* . Observe that this entropy is a non-decreasing function of the partition:

$$\mathcal{P} < \mathcal{Q} \implies h_\mu(f, \mathcal{P}) \leq h_\mu(f, \mathcal{Q}). \tag{9.1.16}$$

Indeed, if  $\mathcal{P} < \mathcal{Q}$  then  $\mathcal{P}^n < \mathcal{Q}^n$  for every  $n$ . Using Lemma 9.1.5, it follows that  $H_\mu(\mathcal{P}^n) \leq H_\mu(\mathcal{Q}^n)$  for every  $n$ , and this implies (9.1.16).

Finally, the *entropy of the system  $(f, \mu)$*  is defined by

$$h_\mu(f) = \sup_{\mathcal{P}} h_\mu(f, \mathcal{P}), \tag{9.1.17}$$

where the supremum is taken over all the partitions with finite entropy. A useful observation is that the definition is not affected if we take the supremum only over the finite partitions (see Exercise 9.1.2).

**Example 9.1.8.** Suppose that the invariant measure  $\mu$  is supported on a periodic orbit. In other words, there exist  $x \in M$  and  $k \geq 1$  such that  $f^k(x) = x$  and the measure  $\mu$  is given by

$$\mu = \frac{1}{k} (\delta_x + \delta_{f(x)} + \dots + \delta_{f^{k-1}(x)}).$$

Note that this measure takes only a finite number of values (because the Dirac measure takes only the values 0 and 1). Hence, the entropy function  $\mathcal{P} \mapsto H_\mu(\mathcal{P})$  also takes only finitely many values. In particular,  $\lim_n \frac{1}{n} H_\mu(\mathcal{P}^n) = 0$  for every partition  $\mathcal{P}$ . This proves that  $h_\mu(f) = 0$ .

**Example 9.1.9.** Consider the decimal expansion map  $f : [0, 1] \rightarrow [0, 1]$ , given by  $f(x) = 10x - [10x]$ . As observed previously,  $f$  preserves the Lebesgue measure  $\mu$  on the interval. Let  $\mathcal{P}$  be the partition of  $[0, 1]$  into the intervals of the form  $((i - 1)/10, i/10]$  with  $i = 1, \dots, 10$ . Then,  $\mathcal{P}^n$  is the partition into the intervals of the form  $((i - 1)/10^n, i/10^n]$  with  $i = 1, \dots, 10^n$ . Using the

calculation in Example 9.1.1, we get that

$$h_\mu(f, \mathcal{P}) = \lim_n \frac{1}{n} H_\mu(\mathcal{P}^n) = \log 10.$$

Using the theory in Section 9.2 (the theorem of Kolmogorov–Sinai and its corollaries), one can easily check that this is also the value of the entropy  $h_\mu(f)$ , that is,  $\mathcal{P}$  realizes the supremum in the definition (9.1.17).

**Example 9.1.10.** Consider the shift map  $\sigma : \Sigma \rightarrow \Sigma$  in  $\Sigma = \{1, \dots, d\}^{\mathbb{N}}$  (or  $\Sigma = \{1, \dots, d\}^{\mathbb{Z}}$ ), with a Bernoulli measure  $\mu = \nu^{\mathbb{N}}$  (respectively,  $\mu = \nu^{\mathbb{Z}}$ ). Let  $\mathcal{P}$  be the partition of  $\Sigma$  into the cylinders  $[0; a]$  with  $a = 1, \dots, d$ . Then,  $\mathcal{P}^n$  is the partition into cylinders  $[0; a_1, \dots, a_n]$  of length  $n$ . Using the calculation in Example 9.1.2 we conclude that

$$h_\mu(\sigma, \mathcal{P}) = \lim_n \frac{1}{n} H_\mu(\mathcal{P}^n) = \sum_{i=1}^d -p_i \log p_i. \tag{9.1.18}$$

The theory presented in Section 9.2 permits us to prove that this is also the value of the entropy  $h_\mu(\sigma)$ .

It follows from expression (9.1.18) that for every  $x > 0$  there exists some Bernoulli shift  $(\sigma, \mu)$  such that  $h_\mu(\sigma) = x$ . We use this observation a few times in what follows.

**Lemma 9.1.11.**  $h_\mu(f, \mathcal{Q}) \leq h_\mu(f, \mathcal{P}) + H_\mu(\mathcal{Q}/\mathcal{P})$  for any partitions  $\mathcal{P}$  and  $\mathcal{Q}$  with finite entropy.

*Proof.* By Lemma 9.1.5, for every  $n \geq 1$ ,

$$\begin{aligned} H_\mu(\mathcal{Q}^{n+1}/\mathcal{P}^{n+1}) &= H_\mu(\mathcal{Q}^n \vee f^{-n}(\mathcal{Q})/\mathcal{P}^n \vee f^{-n}(\mathcal{P})) \\ &\leq H_\mu(\mathcal{Q}^n/\mathcal{P}^n) + H_\mu(f^{-n}(\mathcal{Q})/f^{-n}(\mathcal{P})). \end{aligned}$$

The last term is equal to  $H_\mu(\mathcal{Q}/\mathcal{P})$ , because the measure  $\mu$  is invariant under  $f$ . Therefore, the previous relation proves that

$$H_\mu(\mathcal{Q}^n/\mathcal{P}^n) \leq nH_\mu(\mathcal{Q}/\mathcal{P}) \quad \text{for every } n \geq 1. \tag{9.1.19}$$

Using Lemma 9.1.5 once more, it follows that

$$H_\mu(\mathcal{Q}^n) \leq H_\mu(\mathcal{P}^n \vee \mathcal{Q}^n) = H_\mu(\mathcal{P}^n) + H_\mu(\mathcal{Q}^n/\mathcal{P}^n) \leq H_\mu(\mathcal{P}^n) + nH_\mu(\mathcal{Q}/\mathcal{P}).$$

Dividing by  $n$  and taking the limit when  $n \rightarrow \infty$ , we get the conclusion of the lemma.

**Lemma 9.1.12.**  $h_\mu(f, \mathcal{P}) = \lim_n H_\mu(\mathcal{P}/\bigvee_{j=1}^n f^{-j}(\mathcal{P}))$  for any partition  $\mathcal{P}$  with finite entropy.

*Proof.* Using Lemma 9.1.5(i) and the fact that the measure  $\mu$  is invariant under  $f$ , we get that

$$\begin{aligned} H_\mu\left(\bigvee_{j=0}^{n-1} f^{-j}(\mathcal{P})\right) &= H_\mu\left(\bigvee_{j=1}^{n-1} f^{-j}(\mathcal{P})\right) + H_\mu\left(\mathcal{P}/\bigvee_{j=1}^{n-1} f^{-j}(\mathcal{P})\right) \\ &= H_\mu\left(\bigvee_{j=0}^{n-2} f^{-j}(\mathcal{P})\right) + H_\mu\left(\mathcal{P}/\bigvee_{j=1}^{n-1} f^{-j}(\mathcal{P})\right) \end{aligned}$$

for every  $n$ . By recurrence, it follows that

$$H_\mu\left(\bigvee_{j=0}^{n-1} f^{-j}(\mathcal{P})\right) = H_\mu(\mathcal{P}) + \sum_{k=1}^{n-1} H_\mu\left(\mathcal{P}/\bigvee_{j=1}^k f^{-j}(\mathcal{P})\right).$$

Therefore,  $h_\mu(f, \mathcal{P})$  is given by the Cesàro limit

$$h_\mu(f, \mathcal{P}) = \lim_n \frac{1}{n} H_\mu\left(\bigvee_{j=0}^{n-1} f^{-j}(\mathcal{P})\right) = \lim_n \frac{1}{n} \sum_{k=1}^{n-1} H_\mu\left(\mathcal{P}/\bigvee_{j=1}^k f^{-j}(\mathcal{P})\right).$$

On the other hand, Lemma 9.1.5(ii) ensures that  $H_\mu(\mathcal{P}/\bigvee_{j=1}^n f^{-j}(\mathcal{P}))$  is a non-increasing sequence. In particular,  $\lim_n H_\mu(\mathcal{P}/\bigvee_{j=1}^n f^{-j}(\mathcal{P}))$  exists and, consequently, coincides with the Cesàro limit in the previous identity.

Recall that  $\mathcal{P}^n = \bigvee_{j=0}^{n-1} f^{-j}(\mathcal{P})$ . When  $f : M \rightarrow M$  is invertible, we also consider  $\mathcal{P}^{\pm n} = \bigvee_{j=-n}^{n-1} f^{-j}(\mathcal{P})$ .

**Lemma 9.1.13.** *Let  $\mathcal{P}$  be a partition with finite entropy. For every  $k \geq 1$ , we have  $h_\mu(f, \mathcal{P}) = h_\mu(f, \mathcal{P}^k)$  and, if  $f$  is invertible,  $h_\mu(f, \mathcal{P}) = h_\mu(f, \mathcal{P}^{\pm k})$ .*

*Proof.* Observe that, given any  $n \geq 1$ ,

$$\bigvee_{j=0}^{n-1} f^{-j}(\mathcal{P}^k) = \bigvee_{j=0}^{n-1} f^{-j}\left(\bigvee_{i=0}^{k-1} f^{-i}(\mathcal{P})\right) = \bigvee_{l=0}^{n+k-2} f^{-l}(\mathcal{P}) = \mathcal{P}^{n+k-1}.$$

Therefore,

$$h_\mu(f, \mathcal{P}^k) = \lim_n \frac{1}{n} H_\mu(\mathcal{P}^{n+k-1}) = \lim_n \frac{1}{n} H_\mu(\mathcal{P}^n) = h_\mu(f, \mathcal{P}).$$

This proves the first part of the lemma. To prove the second part, note that

$$\bigvee_{j=0}^{n-1} f^{-j}(\mathcal{P}^{\pm k}) = \bigvee_{j=0}^{n-1} f^{-j}\left(\bigvee_{i=-k}^{k-1} f^{-i}(\mathcal{P})\right) = \bigvee_{l=-k}^{n+k-2} f^{-l}(\mathcal{P}) = f^{-k}(\mathcal{P}^{n+2k-1})$$

for every  $n$  and every  $k$ . Therefore,

$$h_\mu(f, \mathcal{P}^{\pm k}) = \lim_n \frac{1}{n} H_\mu(f^{-k}(\mathcal{P}^{n+2k-1})) = \lim_n \frac{1}{n} H_\mu(\mathcal{P}^{n+2k-1}) = h_\mu(f, \mathcal{P})$$

(the second equality uses the fact that  $\mu$  is invariant under  $f$ ).

**Proposition 9.1.14.** *One has  $h_\mu(f^k) = kh_\mu(f)$  for every  $k \in \mathbb{N}$ . If  $f$  is invertible then  $h_\mu(f^k) = |k|h_\mu(f)$  for every  $k \in \mathbb{Z}$ .*

*Proof.* It is clear that the identity holds for  $k = 0$ , since  $f^0 = \text{id}$  and  $h_\mu(\text{id}) = 0$ . Take  $k$  to be non-zero from now on. Let  $g = f^k$  and  $\mathcal{P}$  be any partition of  $M$  with finite entropy. Recalling that  $\mathcal{P}^k = \mathcal{P} \vee f^{-1}(\mathcal{P}) \vee \dots \vee f^{-k+1}(\mathcal{P})$ , we see that

$$\mathcal{P}^{km} = \bigvee_{j=0}^{km-1} f^{-j}(\mathcal{P}) = \bigvee_{i=0}^{m-1} f^{-ki} \left( \bigvee_{j=0}^{k-1} f^{-j}(\mathcal{P}) \right) = \bigvee_{i=0}^{m-1} g^{-i}(\mathcal{P}^k).$$

Therefore,

$$kh_\mu(f, \mathcal{P}) = \lim_m \frac{1}{m} H_\mu(\mathcal{P}^{km}) = h_\mu(g, \mathcal{P}^k). \tag{9.1.20}$$

Since  $\mathcal{P} \prec \mathcal{P}^k$ , this implies that  $h_\mu(g, \mathcal{P}) \leq kh_\mu(f, \mathcal{P}) \leq h_\mu(g)$  for any  $\mathcal{P}$ . Taking the supremum over these partitions  $\mathcal{P}$ , it follows that  $h_\mu(g) \leq kh_\mu(f) \leq h_\mu(g)$ . This proves that  $kh_\mu(f) = h_\mu(g)$ , as stated.

Now suppose that  $f$  is invertible. Let  $\mathcal{P}$  be any partition of  $M$  with finite entropy. For any  $n \geq 1$ ,

$$H_\mu \left( \bigvee_{j=0}^{n-1} f^{-j}(\mathcal{P}) \right) = H_\mu \left( f^{-n+1} \left( \bigvee_{i=0}^{n-1} f^i(\mathcal{P}) \right) \right) = H_\mu \left( \bigvee_{i=0}^{n-1} f^i(\mathcal{P}) \right),$$

because the measure  $\mu$  is invariant under  $f$ . Dividing by  $n$  and taking the limit when  $n \rightarrow \infty$ , we get that

$$h_\mu(f, \mathcal{P}) = h_\mu(f^{-1}, \mathcal{P}). \tag{9.1.21}$$

Taking the supremum over these partitions  $\mathcal{P}$ , it follows that  $h_\mu(f) = h_\mu(f^{-1})$ . Replacing  $f$  with  $f^k$  and using the first half of the proposition, we get that  $h_\mu(f^{-k}) = h_\mu(f^k) = kh_\mu(f)$  for every  $k \in \mathbb{N}$ .

### 9.1.4 Exercises

- 9.1.1. Prove that  $H_\mu(\mathcal{P}/\mathcal{R}) \leq H_\mu(\mathcal{P}/\mathcal{Q}) + H_\mu(\mathcal{Q}/\mathcal{R})$  for any partitions  $\mathcal{P}$ ,  $\mathcal{Q}$  and  $\mathcal{R}$ .
- 9.1.2. Show that the supremum of  $h_\mu(f, \mathcal{P})$  over the finite partitions coincides with the supremum over all the partitions with finite entropy.
- 9.1.3. Check that  $\lim_n H_\mu(\bigvee_{i=0}^{k-1} f^{-i}(\mathcal{P}) / \bigvee_{j=k}^n f^{-j}(\mathcal{P})) = kh(f, \mathcal{P})$  for every partition  $\mathcal{P}$  with finite entropy and every  $k \geq 1$ .
- 9.1.4. Let  $f : M \rightarrow M$  be a measurable transformation preserving a probability measure  $\mu$ .
  - (a) Assume that there exists an invariant set  $A \subset M$  with  $\mu(A) \in (0, 1)$ . Let  $\mu_A$  and  $\mu_B$  be the normalized restrictions of  $\mu$  to the sets  $A$  and  $B = A^c$ , respectively. Show that  $h_\mu(f) = \mu(A)h_{\mu_A}(f) + \mu(B)h_{\mu_B}(f)$ .
  - (b) Suppose that  $\mu$  is a convex combination  $\mu = \sum_{i=1}^n a_i \mu_i$  of ergodic measures  $\mu_1, \dots, \mu_n$ . Show that  $h_\mu(f) = \sum_{i=1}^n a_i h_{\mu_i}(f)$ .

[Observation: In Section 9.6 we prove much stronger results.]

- 9.1.5. Let  $(M, \mathcal{B}, \mu)$  and  $(N, \mathcal{C}, \nu)$  be probability spaces and  $f : M \rightarrow M$  and  $g : N \rightarrow N$  be measurable transformations preserving the measures  $\mu$  and  $\nu$ , respectively. We

say that  $(g, \nu)$  is a *factor* of  $(f, \mu)$  if there exists a measurable map, not necessarily invertible,  $\phi : (M, \mathcal{B}) \rightarrow (N, \mathcal{C})$  such that  $\phi_*\mu = \nu$  and  $\phi \circ f = g \circ \phi$  at almost every point. Show that in that case  $h_\nu(g) \leq h_\mu(f)$ .

### 9.2 Theorem of Kolmogorov–Sinai

In general, the main difficulty in calculating the entropy lies in the calculation of the supremum in (9.1.17). The methods that we develop in this section permit the simplification of that task in many cases, by identifying certain partitions  $\mathcal{P}$  that realize the supremum, that is, such that  $h_\mu(f, \mathcal{P}) = h_\mu(f)$ . The main result is:

**Theorem 9.2.1** (Kolmogorov–Sinai). *Let  $\mathcal{P}_1 < \dots < \mathcal{P}_n < \dots$  be a non-decreasing sequence of partitions with finite entropy such that  $\bigcup_{n=1}^\infty \mathcal{P}_n$  generates the  $\sigma$ -algebra of measurable sets, up to measure zero. Then,*

$$h_\mu(f) = \lim_n h_\mu(f, \mathcal{P}_n).$$

*Proof.* The limit always exists, for property (9.1.16) implies that the sequence  $h_\mu(f, \mathcal{P}_n)$  is non-decreasing. The inequality  $\geq$  in the statement is a direct consequence of the definition of entropy. Therefore, we only need to show that  $h_\mu(f, \mathcal{Q}) \leq \lim_n h_\mu(f, \mathcal{P}_n)$  for every partition  $\mathcal{Q}$  with finite entropy. We use the following fact, which is interesting in itself:

**Proposition 9.2.2.** *Let  $\mathcal{A}$  be an algebra that generates the  $\sigma$ -algebra of measurable sets, up to measure zero. For every partition  $\mathcal{Q}$  with finite entropy and every  $\varepsilon > 0$  there exists some finite partition  $\mathcal{P} \subset \mathcal{A}$  such that  $H_\mu(\mathcal{Q}/\mathcal{P}) < \varepsilon$ .*

*Proof.* The first step is to reduce the statement to the case when  $\mathcal{Q}$  is finite. Denote by  $Q_j, j = 1, 2, \dots$  the elements of  $\mathcal{Q}$ . For each  $k \geq 1$ , consider the finite partition

$$\mathcal{Q}_k = \left\{ Q_1, \dots, Q_k, M \setminus \bigcup_{j=1}^k Q_j \right\}.$$

**Lemma 9.2.3.** *If  $\mathcal{Q}$  is a partition with finite entropy then  $\lim_k H_\mu(\mathcal{Q}/\mathcal{Q}_k) = 0$ .*

*Proof.* Denote  $Q_0 = M \setminus \bigcup_{j=1}^k Q_j$ . By definition,

$$H_\mu(\mathcal{Q}/\mathcal{Q}_k) = \sum_{i=0}^k \sum_{j \geq 1} -\mu(Q_i \cap Q_j) \log \frac{\mu(Q_i \cap Q_j)}{\mu(Q_i)}.$$

All the terms with  $i \geq 1$  vanish, since in that case  $\mu(Q_i \cap Q_j)$  is equal to zero if  $i \neq j$  and is equal to  $\mu(Q_i)$  if  $i = j$ . For  $i = 0$  we have that  $\mu(Q_i \cap Q_j)$  is equal to zero if  $j \leq k$  and is equal to  $\mu(Q_j)$  if  $j > k$ . Therefore,

$$H_\mu(\mathcal{Q}/\mathcal{Q}_k) = \sum_{j > k} -\mu(Q_j) \log \frac{\mu(Q_j)}{\mu(Q_0)} \leq \sum_{j > k} -\mu(Q_j) \log \mu(Q_j).$$

The hypothesis that  $\mathcal{Q}$  has finite entropy means that the expression on the right-hand side converges to zero when  $k \rightarrow \infty$ .

Given  $\varepsilon > 0$ , fix  $k \geq 1$  such that  $H_\mu(\mathcal{Q}/\mathcal{Q}_k) < \varepsilon/2$ . Consider any  $\delta > 0$ . By the approximation theorem (Theorem A.1.19), for each  $i = 1, \dots, k$  there exists  $A_i \in \mathcal{A}$  such that

$$\mu(Q_i \Delta A_i) < \delta/(2k^2). \tag{9.2.1}$$

Define  $P_1 = A_1$  and  $P_i = A_i \setminus \bigcup_{j=1}^{i-1} A_j$  for  $i = 2, \dots, k$  and  $P_0 = M \setminus \bigcup_{j=1}^{k-1} A_j$ . It is clear that  $\mathcal{P} = \{P_1, \dots, P_k, P_0\}$  is a partition of  $M$  and also that  $P_i \in \mathcal{A}$  for every  $i$ . For  $i = 1, \dots, k$ , we have  $P_i \Delta A_i = P_i \setminus A_i = A_i \cap (\bigcup_{j=1}^{i-1} A_j)$ . For any  $x$  in this set, there is  $j < i$  such that  $x \in A_i \cap A_j$ . Since  $Q_i \cap Q_j = \emptyset$ , it follows that  $x \in (A_i \setminus Q_i) \cup (A_j \setminus Q_j)$ . This proves that

$$P_i \Delta A_i \subset \bigcup_{j=1}^i (A_j \setminus Q_j) \subset \bigcup_{j=1}^i (A_j \Delta Q_j),$$

and so  $\mu(P_i \Delta A_i) < i\delta/(2k^2) \leq \delta/(2k)$ . Together with (9.2.1), this implies that

$$\mu(P_i \Delta Q_i) < \delta/(2k^2) + \delta/(2k) \leq \delta/k \quad \text{for } i = 1, \dots, k. \tag{9.2.2}$$

Moreover,  $P_0 \Delta Q_0 \subset \bigcup_{i=1}^k P_i \Delta Q_i$  since  $\mathcal{P}$  and  $\mathcal{Q}_k$  are partitions of  $M$ . Hence, (9.2.2) implies that

$$\mu(P_0 \Delta Q_0) < \delta. \tag{9.2.3}$$

By Lemma 9.1.6, the relations (9.2.2) and (9.2.3) imply that  $H_\mu(\mathcal{Q}_k/\mathcal{P}) < \varepsilon/2$ , as long as we take  $\delta > 0$  sufficiently small. Then, by the inequality in Exercise 9.1.1,

$$H_\mu(\mathcal{Q}/\mathcal{P}) \leq H_\mu(\mathcal{Q}/\mathcal{Q}_k) + H_\mu(\mathcal{Q}_k/\mathcal{P}) < \varepsilon,$$

as stated.

**Corollary 9.2.4.** *If  $(\mathcal{P}_n)_n$  is a sequence of partitions as in Theorem 9.2.1 then  $\lim_n H_\mu(\mathcal{Q}/\mathcal{P}_n) = 0$  for every partition  $\mathcal{Q}$  with finite entropy.*

*Proof.* For each  $n$ , let  $\mathcal{A}_n$  be the algebra generated by  $\bigcup_{j=1}^n P_j$ . Then  $(\mathcal{A}_n)_n$  is a non-decreasing sequence and the union  $\mathcal{A} = \bigcup_n \mathcal{A}_n$  is the algebra generated by  $\bigcup_n \mathcal{P}_n$ . Consider any  $\varepsilon > 0$ . By Proposition 9.2.2, there exists a finite partition  $\mathcal{P} \subset \mathcal{A}$  such that  $H_\mu(\mathcal{Q}/\mathcal{P}) < \varepsilon$ . Hence, since  $\mathcal{P}$  is finite, there exists  $m \geq 1$  such that  $\mathcal{P} \subset \mathcal{A}_m$  and, thus,  $\mathcal{P}$  is coarser than  $\mathcal{P}_m$ . Then, using Lemma 9.1.5,

$$H_\mu(\mathcal{Q}/\mathcal{P}_n) \leq H_\mu(\mathcal{Q}/\mathcal{P}_m) \leq H_\mu(\mathcal{Q}/\mathcal{P}) < \varepsilon \quad \text{for every } n \geq m.$$

This completes the proof of the corollary.

We are ready to conclude the proof of Theorem 9.2.1. By Lemma 9.1.11,

$$h_\mu(f, \mathcal{Q}) \leq h_\mu(f, \mathcal{P}_n) + H_\mu(\mathcal{Q}/\mathcal{P}_n) \quad \text{for every } n.$$

Taking the limit as  $n \rightarrow \infty$ , we get that  $h_\mu(f, \mathcal{Q}) \leq \lim_n h_\mu(f, \mathcal{P}_n)$  for every partition  $\mathcal{Q}$  with finite entropy.

### 9.2.1 Generating partitions

In this section and the ones that follow, we deduce several useful consequences of Theorem 9.2.1.

**Corollary 9.2.5.** *Let  $\mathcal{P}$  be a partition with finite entropy such that the union of the iterates  $\mathcal{P}^n = \bigvee_{j=0}^{n-1} f^{-j}(\mathcal{P})$ ,  $n \geq 1$  generates the  $\sigma$ -algebra of measurable sets, up to measure zero. Then,  $h_\mu(f) = h_\mu(f, \mathcal{P})$ .*

*Proof.* It suffices to apply Theorem 9.2.1 to the sequence  $\mathcal{P}^n$  and to recall that  $h_\mu(f, \mathcal{P}^n) = h_\mu(f, \mathcal{P})$  for every  $n$ , by Lemma 9.1.13.

**Corollary 9.2.6.** *Assume that the system  $(f, \mu)$  is invertible. Let  $\mathcal{P}$  be a partition with finite entropy such that the union of the iterates  $\mathcal{P}^{\pm n} = \bigvee_{j=-n}^{n-1} f^{-j}(\mathcal{P})$ ,  $n \geq 1$  generates the  $\sigma$ -algebra of measurable sets, up to measure zero. Then,  $h_\mu(f) = h_\mu(f, \mathcal{P})$ .*

*Proof.* It suffices to apply Theorem 9.2.1 to the sequence  $\mathcal{P}^{\pm n}$  and to recall that  $h_\mu(f, \mathcal{P}^{\pm n}) = h_\mu(f, \mathcal{P})$  for every  $n$ , by Lemma 9.1.13.

In particular, Corollaries 9.2.5 and 9.2.6 complete the calculation of the entropy of the decimal expansion and the Bernoulli shifts that we started in Examples 9.1.9 and 9.1.10, respectively.

In both situations, Corollary 9.2.5 and Corollary 9.2.6, we say that  $\mathcal{P}$  is a *generating partition* or, simply, a *generator* of the system. Note, however, that this contains a certain abuse of language, since the conditions in the two corollaries are not equivalent. For example, for the shift map in  $M = \{1, \dots, d\}^{\mathbb{Z}}$ , the partition  $\mathcal{P}$  into cylinders  $\{[0; a] : a = 1, \dots, d\}$  is such that the union of the two-sided iterates  $\mathcal{P}^{\pm n}$  generates the  $\sigma$ -algebra of measurable sets but the union of the one-sided iterates  $\mathcal{P}^n$  does not. Whenever it is necessary to distinguish between these two concepts, we talk of a *one-sided* generator and a *two-sided* generator, respectively.

In this regard, let us point out that certain *invertible* systems admit one-sided generators. For example, if  $f : S^1 \rightarrow S^1$  is an irrational rotation and  $\mathcal{P} = \{I, S^1 \setminus I\}$  is a partition of the circle into two complementary intervals, then  $\mathcal{P}$  is a one-sided generator (and also a two-sided generator, of course). However, this kind of situation is possible only for systems with entropy zero:

**Corollary 9.2.7.** *Assume that the system  $(f, \mu)$  is invertible and there exists a partition  $\mathcal{P}$  with finite entropy such that  $\bigcup_{n=1}^\infty \mathcal{P}^n$  generates the  $\sigma$ -algebra of measurable sets, up to measure zero. Then,  $h_\mu(f) = 0$ .*

*Proof.* Combining Lemma 9.1.12 and Corollary 9.2.5, we get that

$$h_\mu(f) = h_\mu(f, \mathcal{P}) = \lim_n H_\mu(\mathcal{P}/f^{-1}(\mathcal{P}^n)).$$

Since  $\bigcup_n \mathcal{P}^n$  generates the  $\sigma$ -algebra  $\mathcal{B}$  of measurable sets,  $\bigcup_n f^{-1}(\mathcal{P}^n)$  generates the  $\sigma$ -algebra  $f^{-1}(\mathcal{B})$ . Now notice that  $f^{-1}(\mathcal{B}) = \mathcal{B}$ , since  $f$  is invertible. Hence, Corollary 9.2.4 implies that  $H_\mu(\mathcal{P}/f^{-1}(\mathcal{P}^n))$  converges to zero when  $n \rightarrow \infty$ . It follows that  $h_\mu(f) = h_\mu(f, \mathcal{P}) = 0$ .

Now take  $M$  to be a metric space and  $\mu$  to be a Borel probability measure.

**Corollary 9.2.8.** *Let  $\mathcal{P}_1 < \dots < \mathcal{P}_n < \dots$  be a non-decreasing sequence of partitions with finite entropy such that  $\text{diam } \mathcal{P}_n(x) \rightarrow 0$  for  $\mu$ -almost every  $x \in M$ . Then,*

$$h_\mu(f) = \lim_n h_\mu(f, \mathcal{P}_n).$$

*Proof.* Let  $U$  be any non-empty open subset of  $M$ . The hypothesis ensures that for each  $x \in U$  there exists  $n(x)$  such that the set  $P_x = \mathcal{P}_{n(x)}(x)$  is contained in  $U$ . It is clear that  $P_x$  belongs to the algebra  $\mathcal{A}$  generated by  $\bigcup_n \mathcal{P}_n$ . Observe also that  $\mathcal{A}$  is countable, since it consists of the finite unions of elements of the partitions  $\mathcal{P}_n$  together with the complements of such unions. In particular, the map  $x \mapsto P_x$  takes only countably many values. It follows that  $U = \bigcup_{x \in U} P_x$  is in the  $\sigma$ -algebra generated by  $\mathcal{A}$ . This proves that the  $\sigma$ -algebra generated by  $\mathcal{A}$  contains all the open sets and, thus, all the Borel sets. Now, the conclusion follows directly from Theorem 9.2.1.

**Example 9.2.9.** Let  $f : S^1 \rightarrow S^1$  be a homeomorphism and  $\mu$  be any invariant probability measure. Given a finite partition  $\mathcal{P}$  of  $S^1$  into subintervals, denote by  $x_1, \dots, x_m$  their endpoints. For any  $j \geq 1$ , the partition  $f^{-j}(\mathcal{P})$  consists of the subintervals of  $S^1$  determined by the points  $f^{-j}(x_i)$ . This implies that, for each  $n \geq 1$ , the elements of  $\mathcal{P}^n$  have their endpoints in the set

$$\{f^{-j}(x_i) : j = 0, \dots, n - 1 \text{ and } i = 1, \dots, m\}.$$

In particular,  $\#\mathcal{P}^n \leq mn$ . Then, using Lemma 9.1.3,

$$h_\mu(f, \mathcal{P}) = \lim_n \frac{1}{n} H_\mu(\mathcal{P}^n) \leq \lim_n \frac{1}{n} \log \#\mathcal{P}^n = \lim_n \frac{1}{n} \log mn = 0.$$

It follows that  $h_\mu(f) = 0$ : to see that, it suffices to consider any sequence of finite partitions into intervals with diameter going to zero and to apply Corollary 9.2.8.

**Corollary 9.2.10.** *Let  $\mathcal{P}$  be a partition with finite entropy such that we have  $\text{diam } \mathcal{P}^n(x) \rightarrow 0$  for  $\mu$ -almost every  $x \in M$ . Then,  $h_\mu(f) = h_\mu(f, \mathcal{P})$ .*

*Proof.* It suffices to apply Corollary 9.2.8 to the sequence  $\mathcal{P}^n$ , recalling that  $h_\mu(f, \mathcal{P}^n) = h_\mu(f, \mathcal{P})$  for every  $n$ .

Analogously, if  $f$  is invertible and  $\mathcal{P}$  is a partition with finite entropy and such that  $\text{diam } \mathcal{P}^{\pm n}(x) \rightarrow 0$  for  $\mu$ -almost every  $x \in M$ , then  $h_\mu(f) = h_\mu(f, \mathcal{P})$ .



It is known that generators do exist in most interesting cases, although it may be difficult to exhibit a generator explicitly. Indeed, suppose that the ambient  $M$  is a Lebesgue space. Rokhlin [Rok67a, §10] proved that if a system is aperiodic (that is, the set of periodic points has measure zero) and almost every point has a countable (finite or infinite) set of pre-images, then there exists some generator. In particular, every invertible aperiodic system admits some countable generator. In general, this generator may have infinite entropy. But Rokhlin also showed that every invertible aperiodic system with finite entropy admits some two-sided generator with finite entropy. Moreover (Krieger [Kri70]), this generator may be chosen to be finite if the system is ergodic.

### 9.2.2 Semi-continuity of the entropy

Next, we examine some properties of the *entropy function* that associates with each invariant measure  $\mu$  of a given transformation  $f$  the value of the corresponding entropy  $h_\mu(f)$ . We are going to see that this function is usually not continuous. However, under quite broad assumptions, it is *upper semi-continuous*: given any  $\varepsilon > 0$ , one has  $h_\nu(f) \leq h_\mu(f) + \varepsilon$  for every  $\nu$  sufficiently close to  $\mu$ . That holds, in particular, for the class of transformations that we call *expansive*. These facts have important consequences, some of which are explored in Sections 9.2.3 and 9.6. Moreover, we return to this subject in Section 10.5.

Let us start by showing, through an example, that the entropy function may be discontinuous:

**Example 9.2.11.** Let  $f : [0, 1] \rightarrow [0, 1]$  be the decimal expansion map. As we saw in Example 9.1.9, the entropy of  $f$  with respect to the Lebesgue measure  $m$  is  $h_m(f) = \log 10$ . For each  $k \geq 1$ , denote by  $F_k$  the set of fixed points of the iterate  $f^k$ . Observe that  $F_k$  is an invariant set with  $\#F_k = 10^k$ . Observe also that these sets are equidistributed, in the following sense: each interval  $[(i-1)/10^k, i/10^k]$  contains exactly one point of  $F_k$ . Consider the sequence of measures

$$\mu_k = \frac{1}{10^k} \sum_{x \in F_k} \delta_x.$$

The previous observations imply (check!) that each  $\mu_k$  is an invariant probability measure and the sequence  $(\mu_k)_k$  converges to the Lebesgue measure  $m$  in the weak\* topology. Since  $\mu_k$  is supported on a finite set, the same argument as in Example 9.1.8 proves that  $h_{\mu_k}(f) = 0$  for every  $k$ . In particular, we have that  $\lim_k h_{\mu_k}(f) = 0 < h_m(f)$ .

On the other hand, in general, consider any finite partition  $\mathcal{P}$  of  $M$  whose boundary

$$\partial\mathcal{P} = \bigcup_{P \in \mathcal{P}} \partial P$$

satisfies  $\mu(\partial\mathcal{P}) = 0$ . By Theorem 2.1.2 or, more precisely, by the fact that the topology (2.1.5) is equivalent to the weak\* topology, the function  $\nu \mapsto \nu(P)$  is continuous at the point  $\mu$ , for every  $P \in \mathcal{P}$ . Consequently, the function

$$\nu \mapsto H_\nu(\mathcal{P}) = \sum_{P \in \mathcal{P}} -\nu(P) \log \nu(P)$$

is also continuous at  $\mu$ . The hypothesis on  $\mathcal{P}$  also implies that  $\mu(\partial\mathcal{P}^n) = 0$  for every  $n \geq 1$ , since

$$\partial\mathcal{P}^n \subset \partial\mathcal{P} \cup f^{-1}(\partial\mathcal{P}) \cup \dots \cup f^{-n+1}(\partial\mathcal{P}).$$

Thus, the same argument shows that  $\nu \mapsto H_\nu(\mathcal{P}^n)$  is continuous for every  $n$ .

**Proposition 9.2.12.** *Let  $\mathcal{P}$  be a finite partition such that  $\mu(\partial\mathcal{P}) = 0$ . Then the function  $\nu \mapsto h_\nu(f, \mathcal{P})$  is upper semi-continuous at  $\mu$ .*

*Proof.* Recall that, by definition,

$$h_\nu(f, \mathcal{P}) = \inf_n \frac{1}{n} H_\nu(f, \mathcal{P}).$$

It is a well-known easy fact that the infimum of any family of continuous functions is an upper semi-continuous function.

**Corollary 9.2.13.** *Assume that there exists a finite partition  $\mathcal{P}$  such that  $\mu(\partial\mathcal{P}) = 0$  and  $\bigcup_n \mathcal{P}^n$  generates the  $\sigma$ -algebra of measurable sets of  $M$ , up to measure zero. Then the function  $\eta \mapsto h_\eta(f)$  is upper semi-continuous at  $\mu$ .*

*Proof.* By Proposition 9.2.12, given  $\varepsilon > 0$  there exists a neighborhood  $U$  of  $\mu$  such that  $h_\nu(f, \mathcal{P}) \leq h_\mu(f, \mathcal{P}) + \varepsilon$  for every  $\nu \in U$ . By definition,  $h_\mu(f, \mathcal{P}) \leq h_\mu(f)$ . By Corollary 9.2.5, the hypothesis implies that  $h_\nu(f, \mathcal{P}) = h_\nu(f)$  for every  $\nu$ . Therefore,  $h_\nu(f) \leq h_\mu(f) + \varepsilon$  for every  $\nu \in U$ .

Now let us suppose that  $M$  is a compact metric space. As before, take  $\mu$  to be a Borel probability measure. By definition, the diameter  $\text{diam } \mathcal{P}$  of a partition  $\mathcal{P}$  is the supremum of the diameters of its elements. Then we have the following more specialized version of the previous corollary:

**Corollary 9.2.14.** *Assume that there exists  $\varepsilon_0 > 0$  such that every finite partition  $\mathcal{P}$  with  $\text{diam } \mathcal{P} < \varepsilon_0$  satisfies  $\lim_n \text{diam } \mathcal{P}^n = 0$ . Then, the function  $\mu \mapsto$*

$h_\mu(f)$  is upper semi-continuous. Consequently, that function is bounded and its supremum is attained for some measure  $\mu$ .

*Proof.* As we saw in Corollary 9.2.10, the property  $\lim_n \text{diam } \mathcal{P}^n = 0$  implies that  $\bigcup_n \mathcal{P}^n$  generates the  $\sigma$ -algebra of measurable sets. On the other hand, given any invariant probability measure  $\mu$ , it is easy to choose<sup>2</sup> a partition  $\mathcal{P}$  with diameter smaller than  $\varepsilon_0$  and such that  $\mu(\partial\mathcal{P}) = 0$ . It follows from the previous corollary that the entropy function is upper semi-continuous at  $\mu$  and, since  $\mu$  is arbitrary, this gives the first claim in the statement.

The other claims are general consequences of semi-continuity and the fact that the domain of the entropy function, that is, the space  $\mathcal{M}_1(f)$  of all invariant probability measures, is compact.

When  $f$  is invertible we may replace  $\mathcal{P}^n$  by  $\mathcal{P}^{\pm n} = \bigvee_{j=-n}^{n-1} f^{-j}(\mathcal{P})$  in the statement of Corollaries 9.2.13 and 9.2.14. The proof is analogous, using Corollary 9.2.5 and the version of Corollary 9.2.10 for invertible transformations.

### 9.2.3 Expansive transformations

Next, we discuss a rather broad class of transformations that satisfy the conditions in Corollary 9.2.14.

A continuous transformation  $f : M \rightarrow M$  in a metric space is said to be *expansive* if there exists  $\varepsilon_0 > 0$  (called a *constant of expansivity*) such that, given any  $x, y \in M$  with  $x \neq y$ , there exists  $n \in \mathbb{N}$  such that  $d(f^n(x), f^n(y)) \geq \varepsilon_0$ . That is, any two distinct orbits of  $f$  may be distinguished, at a macroscopic scale, at some stage of the iteration.

When  $f$  is invertible, there is also a two-sided version of the notion of expansivity, defined as follows: there exists  $\varepsilon_0 > 0$  such that, given any  $x, y \in M$  with  $x \neq y$  there exists  $n \in \mathbb{Z}$  such that  $d(f^n(x), f^n(y)) \geq \varepsilon_0$ . It is clear that (one-sided) expansive homeomorphisms are also two-sided expansive.

**Example 9.2.15.** Let  $\sigma : \Sigma \rightarrow \Sigma$  be the shift map in  $\Sigma = \{1, \dots, d\}^{\mathbb{N}}$ . Consider in  $\Sigma$  the distance defined by  $d((x_n)_n, (y_n)_n) = 2^{-N}$ , where  $N$  is the smallest value of  $n$  such that  $x_n \neq y_n$ . Note that  $d(\sigma^N(x), \sigma^N(y)) = 2^0 = 1$  if  $x = (x_n)_n$  and  $y = (y_n)_n$  are distinct. This shows that  $\sigma$  is an expansive transformation, with  $\varepsilon_0 = 1$  as a constant of expansivity.

Analogously, the two-sided shift map  $\sigma : \Sigma \rightarrow \Sigma$  in  $\Sigma = \{1, \dots, d\}^{\mathbb{Z}}$  is two-sided expansive (but not one-sided expansive).

<sup>2</sup> For example: for each  $x$  choose  $r_x \in (0, \varepsilon_0)$  such that the boundary of the ball of center  $x$  and radius  $r_x$  has measure zero. Then let  $\mathcal{U}$  be a finite cover of  $M$  by such balls and take for  $\mathcal{P}$  the partition defined by  $\mathcal{U}$ , that is, the partition whose elements are the maximal sets that, for each  $U \in \mathcal{U}$ , are either contained in  $U$  or disjoint from  $U$ ; see Figure 2.1.

We leave it to the reader to check (Exercise 9.2.1) that the decimal expansion transformation  $f(x) = 10x - [10x]$  is also expansive. On the other hand, torus rotations are never expansive.

**Proposition 9.2.16.** *Let  $f : M \rightarrow M$  be an expansive transformation in a compact metric space and let  $\varepsilon_0 > 0$  be a constant of expansivity. Then  $\lim_n \text{diam } \mathcal{P}^n = 0$  for every finite partition  $\mathcal{P}$  with  $\text{diam } \mathcal{P} < \varepsilon_0$ .*

*Proof.* It is clear that the sequence  $\text{diam } \mathcal{P}^n$  is non-increasing. Suppose that its infimum  $\delta$  is positive. Then, for every  $n \geq 1$  there exist points  $x_n$  and  $y_n$  such that  $d(x_n, y_n) > \delta/2$  but  $x_n$  and  $y_n$  belong to the same element of  $\mathcal{P}^n$  and, thus, satisfy

$$d(f^j(x_n), f^j(y_n)) \leq \text{diam } \mathcal{P} < \varepsilon_0 \quad \text{for every } 0 \leq j < n.$$

By compactness, there exists a sequence  $(n_j)_j \rightarrow \infty$  such that  $(x_{n_j})_j$  and  $(y_{n_j})_j$  converge to points  $x$  and  $y$ , respectively. Then,  $d(x, y) \geq \delta/2 > 0$  but  $d(f^j(x), f^j(y)) \leq \text{diam } \mathcal{P} < \varepsilon_0$  for every  $j \geq 0$ . This contradicts the hypothesis that  $\varepsilon_0$  is a constant of expansivity.

**Corollary 9.2.17.** *If  $f : M \rightarrow M$  is an expansive transformation in a compact metric space then the entropy function is upper semi-continuous. Moreover, there exist invariant probability measures  $\mu$  whose entropy  $h_\mu(f)$  is maximum among all the invariant probability measures of  $f$ .*

*Proof.* Combine Proposition 9.2.16 with Corollary 9.2.14.

If the transformation  $f$  is invertible and two-sided expansive, we may replace  $\mathcal{P}^n$  by  $\mathcal{P}^{\pm n}$  in Proposition 9.2.16 and the conclusion of Corollary 9.2.17 also remains valid as stated.

### 9.2.4 Exercises

- 9.2.1. Show that the decimal expansion  $f : [0, 1] \rightarrow [0, 1]$ ,  $f(x) = 10x - [10x]$  is expansive and exhibit a constant of expansivity.
- 9.2.2. Check that for every  $s > 0$  there exists some Bernoulli shift  $(\sigma, \mu)$  whose entropy is equal to  $s$ .
- 9.2.3. Let  $X = \{0\} \cup \{1/n : n \geq 1\}$  and consider the space  $\Sigma = X^{\mathbb{N}}$  endowed with the distance  $d((x_n)_n, (y_n)_n) = 2^{-N} |x_N - y_N|$ , where  $N = \min\{n \in \mathbb{N} : x_n \neq y_n\}$ .
- Verify that the shift map  $\sigma : \Sigma \rightarrow \Sigma$  is not expansive.
  - For each  $k \geq 1$ , let  $\nu_k$  be the probability measure on  $X$  that assigns weight  $1/2$  to each of the points  $1/k$  and  $1/(k+1)$ . Use the Bernoulli measures  $\mu_k = \nu_k^{\mathbb{N}}$  to conclude that the entropy function of the shift is not upper semi-continuous.
  - Let  $\mu$  be the Bernoulli measure associated with any probability vector  $(p_x)_{x \in X}$  such that  $\sum_{x \in X} -p_x \log p_x = \infty$ . Show that  $h_\mu(\sigma)$  is infinite.
- 9.2.4. Let  $f : S^1 \rightarrow S^1$  be a covering map of degree  $d \geq 2$  and  $\mu$  be a probability measure invariant under  $f$ . Show that  $h_\mu(f) \leq \log d$ .

- 9.2.5. Let  $\mathcal{P}$  and  $\mathcal{Q}$  be two partitions with finite entropy. Show that if  $\mathcal{P}$  is coarser than  $\bigvee_{j=0}^{\infty} f^{-j}(\mathcal{Q})$  then  $h_{\mu}(f, \mathcal{P}) \leq h_{\mu}(f, \mathcal{Q})$ .
- 9.2.6. Show that if  $\mathcal{A}$  is an algebra that generates the  $\sigma$ -algebra of measurable sets, up to measure zero, then the supremum of  $h_{\mu}(f, \mathcal{P})$  over the partitions with finite entropy (or even the finite partitions)  $\mathcal{P} \subset \mathcal{A}$  coincides with  $h_{\mu}(f)$ .
- 9.2.7. Consider transformations  $f : M \rightarrow M$  and  $g : N \rightarrow N$  preserving probability measures  $\mu$  and  $\nu$ , respectively. Consider  $f \times g : M \times N \rightarrow M \times N$  defined by  $(f \times g)(x, y) = (f(x), g(y))$ . Show that  $f \times g$  preserves the product measure  $\mu \times \nu$  and that  $h_{\mu \times \nu}(f \times g) = h_{\mu}(f) + h_{\nu}(g)$ .

### 9.3 Local entropy

The theorem of Shannon–McMillan–Breiman, which we discuss in this section, provides a complementary view of the concept of entropy, more detailed and more local in nature. We also mention a topological version of that idea, which is due to Brin–Katok.

**Theorem 9.3.1** (Shannon–McMillan–Breiman). *Given any partition  $\mathcal{P}$  with finite entropy, the limit*

$$h_{\mu}(f, \mathcal{P}, x) = \lim_n -\frac{1}{n} \log \mu(\mathcal{P}^n(x)) \quad \text{exists at } \mu\text{-almost every point.} \quad (9.3.1)$$

*The function  $x \mapsto h_{\mu}(f, \mathcal{P}, x)$  is  $\mu$ -integrable, and the limit in (9.3.1) also holds in  $L^1(\mu)$ . Moreover,*

$$\int h_{\mu}(f, \mathcal{P}, x) d\mu(x) = h_{\mu}(f, \mathcal{P}).$$

*If  $(f, \mu)$  is ergodic then  $h_{\mu}(f, \mathcal{P}, x) = h_{\mu}(f, \mathcal{P})$  at  $\mu$ -almost every point.*

Recall that  $\mathcal{P}^n(x) = \mathcal{P}(x) \cap f^{-1}(\mathcal{P}(f(x))) \cap \dots \cap f^{-n+1}(\mathcal{P}(f^{n-1}(x)))$ , that is,  $\mathcal{P}^n(x)$  is formed by the points whose trajectories remain “close” to the trajectory of  $x$  during  $n$  iterates, in the sense that they visit the same elements of  $\mathcal{P}$ . Theorem 9.3.1 states that the measure of this set has a well-defined exponential rate of decay: at  $\mu$ -almost every point,

$$\mu(\mathcal{P}^n(x)) \approx e^{-nh_{\mu}(f, \mathcal{P}, x)} \quad \text{for every large } n.$$

The proof of the theorem is presented in Section 9.3.1.

The theorem of Brin–Katok that we state in the sequel belongs to the same family of results, but uses a different notion of proximity.

**Definition 9.3.2.** Let  $f : M \rightarrow M$  be a continuous map in a compact metric space. Given  $x \in M$ ,  $n \geq 1$  and  $\varepsilon > 0$ , we call the *dynamical ball* of length  $n$  and radius  $\varepsilon$  around  $x$  the set

$$B(x, n, \varepsilon) = \{y \in M : d(f^j(x), f^j(y)) < \varepsilon \text{ for every } j = 0, 1, \dots, n - 1\}.$$

In other words,

$$B(x, n, \varepsilon) = B(x, \varepsilon) \cap f^{-1}(B(f(x), \varepsilon)) \cap \dots \cap f^{-n+1}(B(f^{n-1}(x), \varepsilon)).$$

Define

$$h_\mu^+(f, \varepsilon, x) = \limsup_n -\frac{1}{n} \log \mu(B(x, n, \varepsilon)) \quad \text{and}$$

$$h_\mu^-(f, \varepsilon, x) = \liminf_n -\frac{1}{n} \log \mu(B(x, n, \varepsilon)).$$

**Theorem 9.3.3** (Brin–Katok). *Let  $\mu$  be a probability measure invariant under  $f$ . The limits*

$$\lim_{\varepsilon \rightarrow 0} h_\mu^+(f, \varepsilon, x) \quad \text{and} \quad \lim_{\varepsilon \rightarrow 0} h_\mu^-(f, \varepsilon, x)$$

*exist and are equal, for  $\mu$ -almost every point. Denoting by  $h_\mu(f, x)$  their common value, the function  $h_\mu(f, \cdot)$  is integrable and*

$$h_\mu(f) = \int h_\mu(f, x) d\mu(x).$$

The proof of this result may be found in the original paper of Brin and Katok [BK83], and is not presented here.

**Example 9.3.4** (Translations in compact groups). Let  $G$  be a compact metrizable group and  $\mu$  be its Haar measure. Every translation of  $G$  has zero entropy with respect to  $\mu$ . Indeed, consider in  $G$  any distance  $d$  that is invariant under all the translations (recall Lemma 6.3.6). Relative to such a distance,

$$L_g^j(B(x, \varepsilon)) = B(L_g^j(x), \varepsilon)$$

for every  $g \in G$  and  $j \in \mathbb{Z}$ . Consequently,  $B(x, n, \varepsilon) = B(x, \varepsilon)$  for every  $n \geq 1$ . Then,

$$h_\mu^\pm(L_g, \varepsilon, x) = \lim_n -\frac{1}{n} \log \mu(B(x, \varepsilon)) = 0$$

for every  $\varepsilon > 0$  and  $x \in G$ . By the theorem of Brin–Katok, it follows that  $h_\mu(L_g) = 0$  for every  $g \in G$ . The same argument applies to every right-translation  $R_g$ .

### 9.3.1 Proof of the Shannon–McMillan–Breiman theorem

Consider the sequence of functions  $\varphi_n : M \rightarrow \mathbb{R}$  defined by

$$\varphi_n(x) = -\log \frac{\mu(\mathcal{P}^n(x))}{\mu(\mathcal{P}^{n-1}(f(x)))}.$$

By telescopic cancellation,

$$-\frac{1}{n} \log \mu(\mathcal{P}^n(x)) = -\frac{1}{n} \log \mu(\mathcal{P}(f^{n-1}(x))) + \frac{1}{n} \sum_{j=0}^{n-2} \varphi_{n-j}(f^j(x)) \quad (9.3.2)$$

for every  $n$  and every  $x$ .

**Lemma 9.3.5.** *The sequence  $-n^{-1} \log \mu(\mathcal{P}(f^{n-1}(x)))$  converges to zero  $\mu$ -almost everywhere and in  $L^1(\mu)$ .*

*Proof.* Start by noting that the function  $x \mapsto -\log \mu(\mathcal{P}(x))$  is integrable:

$$\int |\log \mu(\mathcal{P}(x))| d\mu(x) = \int -\log \mu(\mathcal{P}(x)) d\mu(x) = H_\mu(\mathcal{P}) < \infty.$$

Using Lemma 3.2.5, it follows that  $-(n-1)^{-1} \log \mu(\mathcal{P}(f^{n-1}(x)))$  converges to zero at  $\mu$ -almost every point. Moreover, it is clear that this conclusion is not affected if one replaces  $n-1$  by  $n$  in the denominator. This proves the claim of  $\mu$ -almost everywhere convergence. Next, using the fact that the measure  $\mu$  is invariant and  $H_\mu(\mathcal{P}) < \infty$ ,

$$\left\| -\frac{1}{n} \log \mu(\mathcal{P}(f^{n-1}(x))) \right\|_1 = \frac{1}{n} \int -\log \mu(\mathcal{P}(f^{n-1}(x))) d\mu(x) = \frac{1}{n} H_\mu(\mathcal{P})$$

converges to zero when  $n \rightarrow \infty$ . This proves the convergence in  $L^1(\mu)$ .

Next, we show that the last term in (9.3.2) also converges  $\mu$ -almost everywhere and in  $L^1(\mu)$ .

**Lemma 9.3.6.** *The limit  $\varphi(x) = \lim_n \varphi_n(x)$  exists at  $\mu$ -almost every point.*

*Proof.* For each  $n > 1$ , denote by  $\mathcal{Q}_n$  the partition of  $M$  defined by

$$\mathcal{Q}_n(x) = f^{-1}(\mathcal{P}^{n-1}(f(x))) = f^{-1}(\mathcal{P}(f(x))) \cap \dots \cap f^{-n+1}(\mathcal{P}(f^{n-1}(x))).$$

Note that  $\mu(\mathcal{P}^{n-1}(f(x))) = \mu(\mathcal{Q}_n(x))$  and  $\mathcal{P}^n(x) = \mathcal{P}(x) \cap \mathcal{Q}_n(x)$ . Therefore,

$$\frac{\mu(\mathcal{P}^n(x))}{\mu(\mathcal{P}^{n-1}(f(x)))} = \frac{\mu(\mathcal{P}(x) \cap \mathcal{Q}_n(x))}{\mu(\mathcal{Q}_n(x))}. \tag{9.3.3}$$

For each  $P \in \mathcal{P}$  and  $n > 1$ , consider the conditional expectation (recall Section 5.2.1)

$$e_n(\mathcal{X}_P, x) = \frac{1}{\mu(\mathcal{Q}_n(x))} \int_{\mathcal{Q}_n(x)} \mathcal{X}_P d\mu = \frac{\mu(P \cap \mathcal{Q}_n(x))}{\mu(\mathcal{Q}_n(x))}.$$

Comparing with (9.3.3) we see that

$$e_n(\mathcal{X}_P, x) = \frac{\mu(\mathcal{P}^n(x))}{\mu(\mathcal{P}^{n-1}(f(x)))} \quad \text{for every } x \in P.$$

By Lemma 5.2.1, the limit  $e(\mathcal{X}_P, x) = \lim_n e_n(\mathcal{X}_P, x)$  exists for  $\mu$ -almost every  $x \in M$  and, in particular, for  $\mu$ -almost every  $x \in P$ . Since  $P \in \mathcal{P}$  is arbitrary, this proves that

$$\lim_n \frac{\mu(\mathcal{P}^n(x))}{\mu(\mathcal{P}^{n-1}(f(x)))}$$

exists for  $\mu$ -almost every point. Taking logarithms, we get that  $\lim_n \varphi_n(x)$  exists for  $\mu$ -almost every point, as stated.

**Lemma 9.3.7.** *The function  $\Phi = \sup_n \varphi_n$  is integrable.*

*Proof.* As in the previous lemma, let us consider the partitions  $\mathcal{Q}_n$  defined by  $\mathcal{Q}_n(x) = f^{-1}(\mathcal{P}^{n-1}(f(x)))$ . Fix any  $P \in \mathcal{P}$ . Given  $x \in P$  and  $t > 0$ , it is clear that  $\Phi(x) > t$  if and only if  $\varphi_n(x) > t$  for some  $n$ . Moreover,

$$\varphi_n(x) > t \iff \mu(P \cap \mathcal{Q}_n(x)) < e^{-t} \mu(\mathcal{Q}_n(x))$$

and, in that case,  $\varphi_n(y) > t$  for every  $y \in P \cap \mathcal{Q}_n(x)$ . Therefore, we may write the set  $\{x \in P : \Phi(x) > t\}$  as a disjoint union  $\bigcup_j (P \cap \mathcal{Q}_j)$ , where each  $\mathcal{Q}_j$  belongs to some partition  $\mathcal{Q}_{n(j)}$  and

$$\mu(P \cap \mathcal{Q}_j) < e^{-t} \mu(\mathcal{Q}_j) \quad \text{for every } j.$$

Consequently, for every  $t > 0$  and  $P \in \mathcal{P}$ ,

$$\mu(\{x \in P : \Phi(x) > t\}) = \sum_j \mu(P \cap \mathcal{Q}_j) < e^{-t} \sum_j \mu(\mathcal{Q}_j) \leq e^{-t}. \tag{9.3.4}$$

Then (see Exercise 9.3.1),

$$\begin{aligned} \int \Phi d\mu &= \sum_{P \in \mathcal{P}} \int_P \Phi d\mu = \sum_{P \in \mathcal{P}} \int_0^\infty \mu(\{x \in P : \Phi(x) > t\}) dt \\ &\leq \sum_{P \in \mathcal{P}} \int_0^\infty \min\{e^{-t}, \mu(P)\} dt. \end{aligned}$$

The last integral may be rewritten as follows:

$$\int_0^{-\log \mu(P)} \mu(P) dt + \int_{-\log \mu(P)}^\infty e^{-t} dt = -\mu(P) \log \mu(P) + \mu(P).$$

Combining these two relations:

$$\int \Phi d\mu \leq \sum_{P \in \mathcal{P}} -\mu(P) \log \mu(P) + \mu(P) = H_\mu(\mathcal{P}) + 1 < \infty.$$

This proves the lemma, since  $\Phi$  is non-negative.

**Lemma 9.3.8.** *The function  $\varphi$  is integrable and  $(\varphi_n)_n$  converges to  $\varphi$  in  $L^1(\mu)$ .*

*Proof.* We saw in Lemma 9.3.6 that  $(\varphi_n)_n$  converges to  $\varphi$  at  $\mu$ -almost every point. Since  $0 \leq \varphi_n \leq \Phi$  for every  $n$ , we also have that  $0 \leq \varphi \leq \Phi$ . In particular,  $\varphi$  is integrable. Moreover,  $|\varphi - \varphi_n| \leq \Phi$  for every  $n$  and, thus, we may use the dominated convergence theorem (Theorem A.2.11) to conclude that

$$\lim_n \int |\varphi - \varphi_n| d\mu = \int \lim_n |\varphi - \varphi_n| d\mu = 0.$$

This proves the convergence in  $L^1(\mu)$ .



**Lemma 9.3.9.** *At  $\mu$ -almost every point and in  $L^1(\mu)$ ,*

$$\lim_n \frac{1}{n} \sum_{j=0}^{n-2} \varphi_{n-j}(f^j(x)) = \lim_n \frac{1}{n} \sum_{j=0}^{n-2} \varphi(f^j(x)).$$

*Proof.* By the Birkhoff ergodic theorem (Theorem 3.2.3), the limit on the right-hand side exists at  $\mu$ -almost every point and in  $L^1(\mu)$ , indeed, it is equal to the time average of the function  $\varphi$ . Therefore, it is enough to show that the difference

$$\frac{1}{n} \sum_{j=0}^{n-2} (\varphi_{n-j} - \varphi) \circ f^j \tag{9.3.5}$$

converges to zero at  $\mu$ -almost every point and in  $L^1(\mu)$ . Since the measure  $\mu$  is invariant,  $\|(\varphi_{n-j} - \varphi) \circ f^j\|_1 = \|\varphi_{n-j} - \varphi\|_1$  for every  $j$ . Hence,

$$\left\| \frac{1}{n} \sum_{j=0}^{n-2} (\varphi_{n-j} - \varphi) \circ f^j \right\|_1 \leq \frac{1}{n} \sum_{j=0}^{n-2} \|\varphi_{n-j} - \varphi\|_1.$$

By Lemma 9.3.8 the sequence on the right-hand side converges to zero. This implies that (9.3.5) converges to zero in  $L^1(\mu)$ . We are left to prove that the sequence converges at  $\mu$ -almost every point.

For each fixed  $k \geq 2$ , consider  $\Phi_k = \sup_{i>k} |\varphi_i - \varphi|$ . Note that  $\Phi_k \leq \Phi$  and, thus,  $\Phi_k \in L^1(\mu)$ . Moreover,

$$\begin{aligned} \frac{1}{n} \sum_{j=0}^{n-2} |\varphi_{n-j} - \varphi| \circ f^j &= \frac{1}{n} \sum_{j=0}^{n-k-1} |\varphi_{n-j} - \varphi| \circ f^j + \frac{1}{n} \sum_{j=n-k}^{n-2} |\varphi_{n-j} - \varphi| \circ f^j \\ &\leq \frac{1}{n} \sum_{j=0}^{n-k-1} \Phi_k \circ f^j + \frac{1}{n} \sum_{j=n-k}^{n-2} \Phi \circ f^j. \end{aligned}$$

By the Birkhoff ergodic theorem, the first term on the right-hand side converges to the time average  $\tilde{\Phi}_k$  at  $\mu$ -almost every point. By Lemma 3.2.5, the last term converges to zero at  $\mu$ -almost every point: the lemma implies that  $n^{-1} \Phi \circ f^{n-i}$  converges to zero for any fixed  $i$ . Hence,

$$\limsup_n \frac{1}{n} \sum_{j=0}^{n-2} |\varphi_{n-j} - \varphi|(f^j(x)) \leq \tilde{\Phi}_k(x) \quad \text{at } \mu\text{-almost every point.} \tag{9.3.6}$$

We claim that  $\lim_k \tilde{\Phi}_k(x) = 0$  at  $\mu$ -almost every point. Indeed, the sequence  $(\tilde{\Phi}_k)_k$  is non-increasing and, by Lemma 9.3.6, it converges to zero at  $\mu$ -almost every point. By the monotone convergence theorem (Theorem A.2.9), it follows that  $\int \Phi_k d\mu \rightarrow 0$  when  $k \rightarrow \infty$ . Another consequence is that  $(\tilde{\Phi}_k)_k$  is non-increasing. Hence, using the monotone convergence theorem together

with the Birkhoff ergodic theorem,

$$\int \lim_k \tilde{\Phi}_k d\mu = \lim_k \int \tilde{\Phi}_k d\mu = \lim_k \int \Phi_k d\mu = 0.$$

Since  $\tilde{\Phi}_k$  is non-negative, it follows that  $\lim_k \tilde{\Phi}_k = 0$  at  $\mu$ -almost every point, as we claimed. Therefore, (9.3.6) implies that

$$\lim_n \frac{1}{n} \sum_{j=0}^{n-2} |\varphi_{n-j} - \varphi| \circ f^j = 0$$

at  $\mu$ -almost every point. This completes the proof of the lemma.

It follows from (9.3.2) and Lemmas 9.3.5 and 9.3.9 that

$$h_\mu(f, \mathcal{P}, x) = \lim_n -\frac{1}{n} \log \mu(\mathcal{P}^n(x))$$

exists at  $\mu$ -almost every point and in  $L^1(\mu)$ : in fact, it coincides with the time average  $\tilde{\varphi}(x)$  of the function  $\varphi$ . Then, in particular,

$$\begin{aligned} \int h_\mu(f, \mathcal{P}, x) d\mu(x) &= \lim_n \frac{1}{n} \int -\log \mu(\mathcal{P}^n(x)) d\mu(x) \\ &= \lim_n \frac{1}{n} H_\mu(\mathcal{P}^n) = h_\mu(f, \mathcal{P}). \end{aligned}$$

Moreover, if  $(f, \mu)$  is ergodic then  $h(f, \mathcal{P}, x) = \tilde{\varphi}(x)$  is constant at  $\mu$ -almost every point. That is, in that case  $h_\mu(f, \mathcal{P}, x) = h_\mu(f, \mathcal{P})$  for  $\mu$ -almost every point. This closes the proof of Theorem 9.3.1.

### 9.3.2 Exercises

9.3.1. Check that, for any measurable function  $\varphi : M \rightarrow (0, \infty)$ ,

$$\int \varphi d\mu = \int_0^\infty \mu(\{x \in M : \varphi(x) > t\}) dt.$$

9.3.2. Use Theorem 9.3.1 to calculate the entropy of a Bernoulli shift in  $\Sigma = \{1, \dots, d\}^{\mathbb{N}}$ .

9.3.3. Show that the function  $h_\mu(f, x)$  in Theorem 9.3.3 is  $f$ -invariant. Conclude that if  $(f, \mu)$  is ergodic, then  $h_\mu(f) = h_\mu(f, x)$  for  $\mu$ -almost every  $x$ .

9.3.4. Suppose that  $(f, \mu)$  is ergodic and let  $\mathcal{P}$  be a partition with finite entropy. Show that given  $\varepsilon > 0$  there exists  $k \geq 1$  such that for every  $n \geq k$  there exists  $\mathcal{B}_n \subset \mathcal{P}^n$  such that

$$e^{-n(h_\mu(f, \mathcal{P}) + \varepsilon)} < \mu(B) < e^{-n(h_\mu(f, \mathcal{P}) - \varepsilon)} \quad \text{for every } B \in \mathcal{B}_n,$$

and the measure of the union of the elements of  $\mathcal{B}_n$  is larger than  $1 - \varepsilon$ .

### 9.4 Examples

In this section we illustrate the previous results through a few examples.

#### 9.4.1 Markov shifts

Let  $\Sigma = \{1, \dots, d\}^{\mathbb{N}}$  and  $\sigma : \Sigma \rightarrow \Sigma$  be the shift map. Let  $\mu$  be the Markov measure associated with a stochastic matrix  $P = (P_{i,j})_{i,j}$  and a probability vector  $p = (p_i)_i$ . We are going to prove:

**Proposition 9.4.1.**  $h_\mu(\sigma) = \sum_{a=1}^d p_a \sum_{b=1}^d -P_{a,b} \log P_{a,b}$ .

*Proof.* Consider the partition  $\mathcal{P}$  of  $\Sigma$  into cylinders  $[0; a]$ ,  $a = 1, \dots, d$ . For each  $n$ , the iterate  $\mathcal{P}^n$  is the partition into cylinders  $[0; a_1, \dots, a_n]$  of length  $n$ . Recalling that  $\mu([0; a_1, \dots, a_n]) = p_{a_1} P_{a_1, a_2} \cdots P_{a_{n-1}, a_n}$ , we see that

$$\begin{aligned} H_\mu(\mathcal{P}^n) &= \sum_{a_1, \dots, a_n} -p_{a_1} P_{a_1, a_2} \cdots P_{a_{n-1}, a_n} \log(p_{a_1} P_{a_1, a_2} \cdots P_{a_{n-1}, a_n}) \\ &= \sum_{a_1} -p_{a_1} \log p_{a_1} \sum_{a_2, \dots, a_n} P_{a_1, a_2} \cdots P_{a_{n-1}, a_n} \\ &\quad + \sum_{j=1}^{n-1} \sum_{a_j, a_{j+1}} -\log P_{a_j, a_{j+1}} \sum p_{a_1} P_{a_1, a_2} \cdots P_{a_{n-1}, a_n}, \end{aligned} \tag{9.4.1}$$

where the last sum is over all the values of  $a_1, \dots, a_{j-1}, a_{j+2}, \dots, a_n$ . On the one hand,

$$\sum_{a_2, \dots, a_n} P_{a_1, a_2} \cdots P_{a_{n-1}, a_n} = \sum_{a_n} P_{a_1, a_n}^n = 1,$$

because  $P^n$  is a stochastic matrix. On the other hand,

$$\begin{aligned} \sum p_{a_1} P_{a_1, a_2} \cdots P_{a_{n-1}, a_n} &= \sum_{a_1, a_n} p_{a_1} P_{a_1, a_j}^j P_{a_j, a_{j+1}} P_{a_{j+1}, a_n}^{n-j-1} \\ &= \sum_{a_1} p_{a_1} P_{a_1, a_j}^j P_{a_j, a_{j+1}} = p_{a_j} P_{a_j, a_{j+1}}, \end{aligned}$$

because  $P^{n-j-1}$  is a stochastic matrix and  $pP^j = P^{*j}p = p$ . Replacing these observations in (9.4.1), we get that

$$\begin{aligned} H_\mu(\mathcal{P}^n) &= \sum_{a_1} -p_{a_1} \log p_{a_1} + \sum_{j=1}^{n-1} \sum_{a_j, a_{j+1}} -p_{a_j} P_{a_j, a_{j+1}} \log P_{a_j, a_{j+1}} \\ &= \sum_a -p_a \log p_a + (n-1) \sum_{a,b} -p_a P_{a,b} \log P_{a,b}. \end{aligned}$$

It follows that  $h_\mu(\sigma, \mathcal{P}) = \sum_{a,b} -p_a P_{a,b} \log P_{a,b}$ . Since the family of all cylinders  $[0; a_1, \dots, a_n]$  generates the  $\sigma$ -algebra of  $\Sigma$ , it follows from Corollary 9.2.5 that  $h_\mu(\sigma) = h_\mu(\sigma, \mathcal{P})$ . This completes the proof of theorem.

This conclusion remains valid for two-sided Markov shifts as well, that is, when  $\Sigma = \{1, \dots, d\}^{\mathbb{Z}}$ . The argument is analogous, using Corollary 9.2.6.

### 9.4.2 Gauss map

Now we calculate the entropy of the Gauss map  $G(x) = (1/x) - [1/x]$  relative to the invariant probability measure

$$\mu(E) = \frac{1}{\log 2} \int_E \frac{dx}{1+x}, \tag{9.4.2}$$

which was already studied in Sections 1.3.2 and 4.2.4. The method that we are going to present extends to a much broader class of systems, including the expanding maps of the interval that are defined and discussed in Example 11.1.16.

Let  $\mathcal{P}$  be the partition of  $(0, 1)$  into the subintervals  $(1/(m + 1), 1/m)$  with  $m \geq 1$ . As before, we denote  $\mathcal{P}^n = \bigvee_{j=0}^{n-1} G^{-j}(\mathcal{P})$ . The following facts are used in what follows:

- (A)  $G^n$  maps each  $P_n \in \mathcal{P}^n$  diffeomorphically onto  $(0, 1)$ , for each  $n \geq 1$ .
- (B)  $\text{diam } \mathcal{P}^n \rightarrow 0$  when  $n \rightarrow \infty$ .
- (C) There exists  $C > 1$  such that  $|(G^n)'(y)|/|(G^n)'(x)| \leq C$  for every  $n \geq 1$  and any  $x$  and  $y$  in the same element of the partition  $\mathcal{P}^n$ .
- (D) There exist  $c_1, c_2 > 0$  such that  $c_1 m(P_n) \leq \mu(P_n) \leq c_2 m(P_n)$  for every  $n \geq 1$  and every  $P_n \in \mathcal{P}^n$ , where  $m$  denotes the Lebesgue measure.

It is immediate from the definition that each  $P \in \mathcal{P}$  is mapped by  $G$  diffeomorphically onto  $(0, 1)$ . Property (A) is a consequence, by induction on  $n$ . Using (A) and Lemma 4.2.12, we get that

$$\text{diam } P_n \leq \sup_{x \in P_n} \frac{1}{|(G^n)'(x)|} \leq 2^{-[n/2]}$$

for every  $n \geq 1$  and every  $P_n \in \mathcal{P}^n$ . This implies (B). Property (C) is given by Lemma 4.2.13. Finally, (D) follows directly from (9.4.2).

**Proposition 9.4.2.**  $h_\mu(G) = \int \log |G'| d\mu$ .

*Proof.* Consider the function  $\psi_n(x) = -\log \mu(\mathcal{P}^n(x))$ , for each  $n \geq 1$ . Observe that

$$H_\mu(\mathcal{P}^n) = \sum_{P_n \in \mathcal{P}^n} -\mu(P_n) \log \mu(P_n) = \int \psi_n(x) d\mu(x).$$

Property (D) gives that

$$-\log c_1 \geq \psi_n(x) + \log m(\mathcal{P}^n(x)) \geq -\log c_2.$$

By property (A), we have  $\log m(\mathcal{P}^n(x)) = -\log |(G^n)'(y)|$  for some  $y \in \mathcal{P}^n(x)$ . Using property (C), it follows that

$$-\log c_1 + \log C \geq \psi_n(x) - \log |(G^n)'(x)| \geq -\log c_2 - \log C$$

for every  $x$  and every  $n$ . Consequently,

$$-\log(c_1/C) \geq H_\mu(\mathcal{P}^n) - \int \log |(G^n)'| d\mu \geq \log(c_2C) \tag{9.4.3}$$

for every  $n$ . Since the measure  $\mu$  is invariant under  $G$ ,

$$\int \log |(G^n)'| d\mu = \sum_{j=0}^{n-1} \int \log |G'| \circ G^j d\mu = n \int |G'| d\mu.$$

Then, dividing (9.4.3) by  $n$  and taking the limit when  $n \rightarrow \infty$ ,

$$h_\mu(f, \mathcal{P}) = \lim_n \frac{1}{n} H_\mu(\mathcal{P}^n) = \int \log |G'| d\mu.$$

Now, property (B) ensures that we may apply Corollary 9.2.10 to conclude that

$$h_\mu(G) = h_\mu(G, \mathcal{P}) = \int \log |G'| d\mu.$$

This completes the proof of the proposition.

The integral in the statement of the proposition may be calculated explicitly: we leave it to the reader to check (using integration by parts and the fact that  $\sum_{j=1}^\infty 1/j^2 = \pi^2/6$ ) that

$$h_\mu(G) = \int \log |G'| d\mu = \int_0^1 \frac{-2 \log x dx}{(1+x) \log 2} = \frac{\pi^2}{6 \log 2} \approx 5.46 \dots$$

Then, recalling that  $(G, \mu)$  is ergodic (Section 4.2.4), it follows from the theorem of Shannon–McMillan–Breiman (Theorem 9.3.1) that

$$\lim_n -\frac{1}{n} \log \mu(\mathcal{P}^n(x)) = \frac{\pi^2}{6 \log 2} \quad \text{for } \mu\text{-almost every } x.$$

As the measure  $\mu$  is comparable to the Lebesgue measure, up to a constant factor, this means that

$$\text{diam } \mathcal{P}^n(x) \approx e^{-\frac{\pi^2 n}{6 \log 2}} \quad (\text{up to a factor } e^{\pm \epsilon n})$$

for  $\mu$ -almost every  $x$  and every  $n$  sufficiently large. Observe that  $\mathcal{P}^n(x)$  is formed by the points  $y$  whose continued fraction expansion coincide with the continued fraction expansion of  $x$  up to order  $n$ .

### 9.4.3 Linear endomorphisms of the torus

Given a real number  $x > 0$ , we denote  $\log^+ x = \max\{\log x, 0\}$ . In this section we prove the following result:

**Proposition 9.4.3.** *Let  $f_A : \mathbb{T}^d \rightarrow \mathbb{T}^d$  be the endomorphism induced on the torus  $\mathbb{T}^d$  by some invertible matrix  $A$  with integer coefficients. Let  $\mu$  be the*

Haar measure of  $\mathbb{T}^d$ . Then

$$h_\mu(f_A) = \sum_{i=1}^d \log^+ |\lambda_i|,$$

where  $\lambda_1, \dots, \lambda_d$  are the eigenvalues of  $A$ , counted with multiplicity.

Initially, let us suppose that the matrix  $A$  is diagonalizable. Let  $v_1, \dots, v_d$  be a normed basis of  $\mathbb{R}^d$  such that  $Av_i = \lambda_i v_i$  for each  $i$ . Let  $u$  be the number of eigenvalues of  $A$  with absolute value strictly larger than 1. We may take the eigenvalues to be numbered in such a way that  $|\lambda_i| > 1$  if and only if  $i \leq u$ . Given  $x \in \mathbb{T}^d$ , every point  $y$  in a neighborhood of  $x$  may be written in the form

$$y = x + \sum_{i=1}^d t_i v_i$$

with  $t_1, \dots, t_d$  close to zero. Given  $\varepsilon > 0$ , denote by  $D(x, \varepsilon)$  the set of points  $y$  of this form with  $|t_i| < \varepsilon$  for every  $i = 1, \dots, d$ . Moreover, for each  $n \geq 1$ , consider

$$D(x, n, \varepsilon) = \{y \in \mathbb{T}^d : f_A^j(y) \in D(f_A^j(x), \varepsilon) \text{ for every } j = 0, \dots, n-1\}.$$

Observe that  $f_A^j(y) = f_A^j(x) + \sum_{i=1}^d t_i \lambda_i^j v_i$  for every  $n \geq 1$ . Therefore,

$$D(x, n, \varepsilon) = \left\{ x + \sum_{i=1}^d t_i v_i : |\lambda_i^n t_i| < \varepsilon \text{ for } i \leq u \text{ and } |t_i| < \varepsilon \text{ for } i > u \right\}.$$

Hence, there exists a constant  $C_1 > 1$  that depends only on  $A$ , such that

$$C_1^{-1} \varepsilon^d \prod_{i=1}^u |\lambda_i|^{-n} \leq \mu(D(x, n, \varepsilon)) \leq C_1 \varepsilon^d \prod_{i=1}^u |\lambda_i|^{-n}$$

for every  $x \in \mathbb{T}^d$ ,  $n \geq 1$  and  $\varepsilon > 0$ . It is also clear that there exists a constant  $C_2 > 1$  that depends only on  $A$ , such that

$$B(x, C_2^{-1} \varepsilon) \subset D(x, \varepsilon) \subset B(x, C_2 \varepsilon)$$

for  $x \in \mathbb{T}^d$  and  $\varepsilon > 0$  small. Then,  $B(x, n, C_2^{-1} \varepsilon) \subset D(x, n, \varepsilon) \subset B(x, n, C_2 \varepsilon)$  for every  $n \geq 1$ . Combining these two observations and taking  $C = C_1 C_2^d$ , we get that

$$C^{-1} \varepsilon^d \prod_{i=1}^u |\lambda_i|^{-n} \leq \mu(B(x, n, \varepsilon)) \leq C \varepsilon^d \prod_{i=1}^u |\lambda_i|^{-n}$$

for every  $x \in \mathbb{T}^d$ ,  $n \geq 1$  and  $\varepsilon > 0$ . Then,

$$h_\mu^+(f, \varepsilon, x) = h_\mu^-(f, \varepsilon, x) = \lim_n \frac{1}{n} \log \mu(B(x, n, \varepsilon)) = \sum_{i=1}^u \log |\lambda_i|$$

for  $x \in \mathbb{T}$  and  $\varepsilon > 0$  small. Hence, using the theorem of Brin–Katok (Theorem 9.3.3),

$$h_\mu(f) = h_\mu(f, x) = \sum_{i=1}^u \log |\lambda_i|$$

for  $\mu$ -almost every point  $x$ . This proves Proposition 9.4.3 in the diagonalizable case.

The general case may be treated analogously, through writing the matrix  $A$  in canonical Jordan form. We leave this task to the reader (Exercise 9.4.2).

### 9.4.4 Differentiable maps

Here we take  $M$  to be a Riemannian manifold (check Appendix A.4.5) and  $f : M \rightarrow M$  to be a *local diffeomorphism*, that is, a  $C^1$  map whose derivative  $Df(x) : T_x M \rightarrow T_{f(x)} M$  at each point  $x$  is an isomorphism. We are going to state and discuss two important theorems, the *Margulis–Ruelle inequality* and the *Pesin entropy formula*, that relate the entropy  $h_\mu(f)$  of an invariant measure to the Lyapunov exponents of the derivative  $Df$ .

Let  $\mu$  be any probability measure invariant under  $f$ . According to the multiplicative ergodic theorem of Oseledets (see Section 3.3.5), for  $\mu$ -almost every point  $x \in M$  there exist  $k(x) \geq 1$ , real numbers  $\lambda_1(x) > \dots > \lambda_k(x)$  and a filtration  $\mathbb{R}^d = V_x^1 > \dots > V_x^{k(x)} > V_x^{k(x)+1} = \{0\}$  such that

$$Df(x)V_x^i = V_{f(x)}^i \quad \text{and} \quad \lim_n \frac{1}{n} \log \|Df^n(x)v\| = \lambda_i(x)$$

for every  $v \in V_x^i \setminus V_x^{i+1}$ , every  $i \in \{1, \dots, k(x)\}$  and  $\mu$ -almost every  $x \in M$ . Moreover, all these objects depend measurably on the point  $x \in M$ . When the measure  $\mu$  is ergodic, the number  $k(x)$ , the Lyapunov exponents  $\lambda_i(x)$  and their multiplicities  $m_i(x) = \dim V_x^i - \dim V_x^{i+1}$  are all constant on a full measure set.

Define  $\rho^+(x)$  to be the sum of all positive Lyapunov exponents, counted with multiplicity:

$$\rho^+(x) = \sum_{i=1}^{k(x)} m_i(x)\lambda_i^+(x) \quad \text{with} \quad \lambda_i^+(x) = \max\{\lambda_i(x), 0\}.$$

The Margulis–Ruelle inequality asserts that the average of  $\rho^+$  is always an upper bound for the entropy of  $(f, \mu)$ . Proofs can be found in Ruelle [Rue78] and Mañé [Mañ87, Section 4.12].

**Theorem 9.4.4** (Margulis–Ruelle inequality).

$$h_\mu(f) \leq \int \rho^+ d\mu.$$

It may happen that all the Lyapunov exponents are positive: that is the case, for instance, for the expanding differentiable maps in Section 11.1. Then,  $\rho^+(x)$  is simply the sum of all Lyapunov exponents, counted with multiplicity. Now, it is also part of the theorem of Oseledets (property (c1) in Section 3.3.5) that

$$\sum_{i=1}^{k(x)} m_i(x)\lambda_i(x) = \lim_n \frac{1}{n} \log |\det Df^n(x)|$$

at  $\mu$ -almost every point. Observe that the right-hand side of this identity is a Birkhoff time average:

$$\lim_n \frac{1}{n} \log |\det Df^n(x)| = \lim_n \frac{1}{n} \sum_{j=0}^{n-1} \log |\det Df|(f^j(x)).$$

So, by the Birkhoff ergodic theorem, the integral of  $\rho^+$  coincides with the integral of the function  $\log |\det Df|$ . Thus, in this case the Margulis–Ruelle inequality becomes:

$$h_\mu(f) \leq \int \log |\det Df| d\mu.$$

Another interesting particular case is when  $f$  is a diffeomorphism. It follows from the version of the Oseledets theorem for invertible maps (also stated in Section 3.3.5) that the Lyapunov exponents of  $Df^{-1}$  are the numbers  $-\lambda_i(x)$ , with multiplicities  $m_i(x)$ . Then, applying Theorem 9.4.4 to the inverse and recalling (Proposition 9.1.14) that  $h_\mu(f) = h_\mu(f^{-1})$ , we get that

$$h_\mu(f) \leq \int \rho^- d\mu, \tag{9.4.4}$$

where  $\rho^-(x) = \sum_{i=1}^{k(x)} m_i(x) \max\{-\lambda_i(x), 0\}$ .

Now let us suppose that the invariant measure  $\mu$  is absolutely continuous with respect to the volume measure associated with the Riemannian structure of  $M$  (check Appendix A.4.5). In this case, assuming just a little bit more regularity, we have a much stronger result:

**Theorem 9.4.5** (Pesin entropy formula). *Assume that the derivative  $Df$  is Hölder and the invariant measure  $\mu$  is absolutely continuous. Then*

$$h_\mu(f) = \int \rho^+ d\mu.$$

This fundamental result was originally proven by Pesin [Pes77]. See also Mañé [Mañ87, Section 4.13] for an alternative proof.

The expression for the entropy of the Haar measure of a linear torus endomorphism, given in Proposition 9.4.3, is a special case of Theorem 9.4.5. Indeed, one can check that the Lyapunov exponents of a linear endomorphism  $f_A$  at every point coincide with the logarithms  $\log |\lambda_i|$  of the absolute values of the eigenvalues of the matrix  $A$ , with the same multiplicities. Thus, in this context

$$\rho^+(x) \equiv \sum_{i=1}^d \log^+ |\lambda_i|.$$

Of course, the Haar measure is absolutely continuous. Another special case of the Pesin entropy formula will appear in Section 12.1.8: see (12.1.31)–(12.1.36).



Finally, let us point out that the assumption of absolute continuity is too strong: the conclusion of Theorem 9.4.5 still holds if the invariant measure is just “absolutely continuous along unstable manifolds”. Roughly speaking, this technical condition means that the conditional probabilities of  $\mu$  with respect to a certain measurable partition, whose elements are unstable disks,<sup>3</sup> are absolutely continuous with respect to the volume measures induced on each of the disks by the Riemannian metric of  $M$ . Moreover, and most striking, this sufficient condition is also necessary for the Pesin entropy formula to hold. For precise statements, related results and proofs, see [LS82, Led84, LY85a, LY85b].

### 9.4.5 Exercises

- 9.4.1. Show that every rotation  $R_\theta : \mathbb{T}^d \rightarrow \mathbb{T}^d$  has entropy zero with respect to the Haar measure of the torus  $\mathbb{T}^d$ . [Observation: This is a special case of Example 9.3.4 but for the present statement we do not need the theorem of Brin–Katok.]
- 9.4.2. Complete the proof of Proposition 9.4.3.
- 9.4.3. Let  $f : M \rightarrow M$  be a measurable transformation and  $\mu$  be an ergodic probability measure. Let  $B \subset M$  be a measurable set with  $\mu(B) > 0$ ,  $g : B \rightarrow B$  be the first-return map of  $f$  to  $B$  and  $\nu$  be the normalized restriction of  $\mu$  to the set  $B$  (recall Section 1.4.1). Show that  $h_\mu(f) = \nu(B)h_\nu(g)$ .
- 9.4.4. Let  $f : M \rightarrow M$  be a measure-preserving transformation in a Lebesgue space  $(M, \mu)$ . Let  $\hat{f} : \hat{M} \rightarrow \hat{M}$  be the natural extension of  $f$  and  $\hat{\mu}$  be the lift of  $\mu$  (Exercise 8.5.7). Show that  $h_\mu(f) = h_{\hat{\mu}}(\hat{f})$ .
- 9.4.5. Prove that if  $f$  is the time-1 of a smooth flow on a surface  $M$  then  $h_\mu(f) = 0$  for every invariant ergodic measure  $\mu$ . [Observation: Using Theorem 9.6.2 below, it follows that the entropy is zero for every invariant measure.]

## 9.5 Entropy and equivalence

The notion of entropy was originally introduced in ergodic theory as a means to distinguish systems that are not ergodically equivalent, especially in the case of systems that are spectrally equivalent and, thus, cannot be distinguished by spectral invariants. It is easy to see that the entropy is, indeed, an invariant of ergodic equivalence:

**Proposition 9.5.1.** *Let  $f : M \rightarrow M$  and  $g : N \rightarrow N$  be transformations preserving probability measures  $\mu$  in  $M$  and  $\nu$  in  $N$ , respectively. If  $(f, \mu)$  is ergodically equivalent to  $(g, \nu)$ , then  $h_\mu(f) = h_\nu(g)$ .*

<sup>3</sup> Unstable disks are differentially embedded disks that are contracted exponentially under negative iteration; in the non-invertible case, the definition is formulated in terms of the natural extension of the transformation.

*Proof.* Let  $\phi : M \rightarrow N$  be an ergodic equivalence between the two systems. This means that  $\phi_*\mu = \nu$  and there exist full measure subsets  $X \subset M$  and  $Y \subset N$  such that  $\phi$  is a measurable bijection from  $X$  to  $Y$ , with measurable inverse. Moreover, as observed in Section 8.1, the sets  $X$  and  $Y$  may be chosen to be invariant. Given any partition  $\mathcal{P}$  of  $M$  with finite entropy for  $\mu$ , let  $\mathcal{P}_X$  be its restriction to  $X$  and  $\mathcal{Q}_Y = \phi(\mathcal{P}_X)$  be the image of  $\mathcal{P}_X$  under  $\phi$ . Then  $\mathcal{Q} = \mathcal{Q}_Y \cup \{Y^c\}$  is a partition of  $N$  and, since  $X$  and  $Y$  are full measure subsets,

$$H_\nu(\mathcal{Q}) = \sum_{Q \in \mathcal{Q}_Y} -\nu(Q) \log \nu(Q) = \sum_{P \in \mathcal{P}_X} -\mu(P) \log \mu(P) = H_\mu(\mathcal{P}).$$

Since  $X$  and  $Y$  are both invariant,  $\mathcal{P}_X^n$  is the restriction of  $\mathcal{P}^n$  to the subset  $X$  and  $\mathcal{Q}^n = \mathcal{Q}_Y^n \cup \{Y^c\}$  for every  $n$ . Moreover,

$$\mathcal{Q}_Y^n = \bigvee_{j=0}^{n-1} g^{-j}(\mathcal{Q}_Y) = \phi \left( \bigvee_{j=0}^{n-1} f^{-j}(\mathcal{P}_X) \right) = \phi(\mathcal{P}_X^n)$$

for every  $n$ . Thus, the previous argument proves that  $H_\nu(\mathcal{Q}^n) = H_\mu(\mathcal{P}^n)$  for every  $n$  and so

$$h_\nu(g, \mathcal{Q}) = \lim_n \frac{1}{n} H_\nu(\mathcal{Q}^n) = \lim_n \frac{1}{n} H_\mu(\mathcal{P}^n) = h_\mu(f, \mathcal{P}).$$

Taking the supremum over  $\mathcal{P}$ , we conclude that  $h_\nu(g) \geq h_\mu(f)$ . The converse inequality is entirely analogous.

Using this observation, Kolmogorov and Sinai concluded that not all two-sided Bernoulli shifts are ergodically equivalent despite their being spectrally equivalent, as we saw in Corollary 8.4.12. This also shows that spectral equivalence is strictly weaker than ergodic equivalence. In fact, as observed in Exercise 9.2.2, for every  $s > 0$  there exists some two-sided Bernoulli shift  $(\sigma, \mu)$  such that  $h_\mu(\sigma) = s$ . Therefore, a sole class of spectral equivalence may contain a whole continuum of ergodic equivalence classes.

### 9.5.1 Bernoulli automorphisms

The converse to Proposition 9.5.1 is false, in general. Indeed, we saw in Example 9.2.9 (and Corollary 9.2.7) that all the circle rotations have entropy zero. But an irrational rotation is never ergodically equivalent to a rational rotation, since the former is ergodic and the latter is not. Besides, Corollary 8.3.6 shows that irrational rotations are also not ergodically equivalent to each other, in general. The case of rational rotations is treated in Exercise 8.3.3.

However, a remarkable result due to Donald Ornstein [Orn70] states that the entropy is a complete invariant for two-sided Bernoulli shifts:

**Theorem 9.5.2 (Ornstein).** *Two-sided Bernoulli shifts in Lebesgue spaces are ergodically equivalent if and only if they have the same entropy.*

We call *Bernoulli automorphism* any system that is ergodically equivalent to a two-sided Bernoulli shift. In the sequel we find several examples of such systems. The theorem of Ornstein may be reformulated as follows: two Bernoulli automorphisms in Lebesgue spaces are ergodically equivalent if and only if they have the same entropy.

Let us point out that the theorem of Ornstein does not extend to one-sided Bernoulli shifts. Indeed, Exercise 8.1.2 shows that in the non-invertible case there are other invariants of ergodic equivalence, including the degree of the transformation (the number of pre-images).

William Parry and Peter Walters [PW72b, PW72a, Wal73] proved, among other results, that one-sided Bernoulli shifts corresponding to probability vectors  $p = (p_1, \dots, p_k)$  and  $q = (q_1, \dots, q_l)$  are ergodically equivalent if and only if  $k = l$  and the vector  $p$  is a permutation of the vector  $q$ . In Exercise 9.7.7, after introducing the notion of the Jacobian, we invite the reader to prove this fact.

### 9.5.2 Systems with entropy zero

In this section we study some properties of systems whose entropy is equal to zero. The main result (Proposition 9.5.5) is that such systems are invertible at almost every point, if the ambient space is a Lebesgue space. It is worthwhile comparing this statement with Corollary 9.2.7. At the end of the section we briefly discuss the spectral types of systems with entropy zero.

In what follows,  $(M, \mathcal{B}, \mu)$  is a probability space and  $f : M \rightarrow M$  is a measure-preserving transformation. In the second half of the section we take  $(M, \mathcal{B}, \mu)$  to be a Lebesgue space.

**Lemma 9.5.3.** *For every  $\varepsilon > 0$  there exists  $\delta > 0$  such that if  $\mathcal{P}$  and  $\mathcal{Q}$  are partitions with finite entropy and  $H_\mu(\mathcal{P}/\mathcal{Q}) < \delta$  then for every  $P \in \mathcal{P}$  there exists a union  $P'$  of elements of  $\mathcal{Q}$  satisfying  $\mu(P \Delta P') < \varepsilon$ .*

*Proof.* Let  $s = 1 - \varepsilon/2$  and  $\delta = -(\varepsilon/2) \log s$ . For each  $P \in \mathcal{P}$  consider

$$\mathcal{S} = \{Q \in \mathcal{Q} : \mu(P \cap Q) \geq s\mu(Q)\}.$$

Let  $P'$  be the union of all the elements of  $\mathcal{S}$ . On the one hand,

$$\mu(P' \setminus P) = \sum_{Q \in \mathcal{S}} \mu(Q \setminus P) \leq \sum_{Q \in \mathcal{S}} (1 - s)\mu(Q) \leq \frac{\varepsilon}{2}. \tag{9.5.1}$$

On the other hand,

$$\begin{aligned} H_\mu(\mathcal{P}/\mathcal{Q}) &= \sum_{R \in \mathcal{P}} \sum_{Q \in \mathcal{Q}} -\mu(R \cap Q) \log \frac{\mu(R \cap Q)}{\mu(Q)} \\ &\geq \sum_{Q \notin \mathcal{S}} -\mu(P \cap Q) \log s = -\mu(P \setminus P') \log s. \end{aligned}$$

This implies that

$$\mu(P \setminus P') \leq \frac{H_\mu(\mathcal{P}/\mathcal{Q})}{-\log s} < \frac{\delta}{-\log s} = \frac{\varepsilon}{2}. \tag{9.5.2}$$

Putting (9.5.1) and (9.5.2) together, we get the conclusion of the lemma.

The next lemma means that the rate  $h_\mu(f, \mathcal{P})$  of information (relative to the partition  $\mathcal{P}$ ) generated by the system at each iteration is zero if and only if *the future determines the present*, in the sense that information relative to the 0-th iterate may be deduced from the ensemble of information relative to the future iterates.

**Lemma 9.5.4.** *Let  $\mathcal{P}$  be a partition with finite entropy. Then,  $h_\mu(f, \mathcal{P}) = 0$  if and only if  $\mathcal{P} \prec \bigvee_{j=1}^\infty f^{-j}(\mathcal{P})$ .*

*Proof.* Suppose that  $h_\mu(f, \mathcal{P})$  is zero. Using Lemma 9.1.12, we obtain that  $H_\mu(\mathcal{P}/\bigvee_{j=1}^n f^{-j}(\mathcal{P}))$  converges to zero when  $n \rightarrow \infty$ . Then, by Lemma 9.5.3, for each  $l \geq 1$  and each  $P \in \mathcal{P}$  there exist  $n_l \geq 1$  and a union  $P_l$  of elements of the partition  $\bigvee_{j=1}^{n_l} f^{-j}(\mathcal{P})$  such that  $\mu(P \Delta P_l) < 2^{-l}$ . It is clear that every  $P_l$  is a union of elements of  $\bigvee_{j=1}^\infty f^{-j}(\mathcal{P})$  and, thus, the same is true for every  $\bigcup_{l=n}^\infty P_l$  and also for  $P_* = \bigcap_{n=1}^\infty \bigcup_{l=n}^\infty P_l$ . Moreover,

$$\mu\left(P \setminus \bigcup_{l=n}^\infty P_l\right) = 0 \quad \text{and} \quad \mu\left(\bigcup_{l=n}^\infty P_l \setminus P\right) \leq 2^{-n}$$

for every  $n$  and, consequently,  $\mu(P \Delta P_*) = 0$ . This shows that every element of  $\mathcal{P}$  coincides, up to measure zero, with a union of elements of  $\bigvee_{j=1}^\infty f^{-j}(\mathcal{P})$ , as claimed in the “only if” half of the statement.

The argument to prove the converse is similar to the one in Proposition 9.2.2. Suppose that  $\mathcal{P} \prec \bigvee_{j=1}^\infty f^{-j}(\mathcal{P})$ . Write  $\mathcal{P} = \{P_j : j = 1, 2, \dots\}$  and, for each  $k \geq 1$ , consider the finite partition  $\mathcal{P}_k = \{P_1, \dots, P_k, M \setminus \bigcup_{j=1}^k P_j\}$ . Given any  $\varepsilon > 0$ , Lemma 9.2.3 ensures that  $H_\mu(\mathcal{P}/\mathcal{P}_k) < \varepsilon/2$  for every  $k$  sufficiently large. Fix  $k$  in these conditions and write  $P_0 = M \setminus \bigcup_{j=1}^k P_j$ . For each  $n \geq 1$  and each  $j = 1, \dots, k$ , let  $Q_j^n$  be the union of the elements of  $\bigvee_{j=1}^n f^{-j}(\mathcal{P})$  that intersect  $P_j$ . The hypothesis ensures that each  $(Q_j^n)_n$  is a decreasing sequence whose intersection coincides with  $P_j$  up to measure zero. Then, given  $\delta > 0$  there exists  $n_0$  such that

$$\sum_{i=1}^k \mu(Q_i^n \setminus P_i) < \delta \quad \text{for every } n \geq n_0. \tag{9.5.3}$$

Define  $R_i^n = Q_i^n \setminus \bigcup_{j=1}^{i-1} Q_j^n$  for  $i = 1, \dots, k$  and also  $R_0^n = M \setminus \bigcup_{j=1}^k Q_j^n$ . It is clear from the construction that  $\mathcal{R}_n = \{R_1^n, \dots, R_k^n, R_0^n\}$  is a partition of  $M$  coarser than  $\bigvee_{j=1}^n f^{-j}(\mathcal{P})$ . Since  $R_i^n \subset Q_i^n$  and  $P_i \subset Q_i^n$ , and the elements of  $\mathcal{P}$  are pairwise disjoint,

$$R_i^n \setminus P_i \subset Q_i^n \setminus P_i \quad \text{and} \quad P_i \setminus R_i^n = P_i \cap \left(\bigcup_{j=1}^{i-1} Q_j^n\right) \subset \bigcup_{j=1}^{i-1} (Q_j^n \setminus P_j)$$

for  $i = 1, \dots, k$ . Similarly,  $R_0^n \subset P_0$  and  $P_0 \setminus R_0^n = P_0 \cap \bigcup_{j=1}^k Q_j^n = \bigcup_{j=1}^k (Q_j^n \setminus P_j)$ . Therefore, the relation (9.5.3) implies that

$$\mu(P_i \Delta R_i^n) < \delta \quad \text{for every } i = 0, 1, \dots, k \text{ and every } n \geq n_0.$$

Then, assuming that  $\delta > 0$  is small, it follows from Lemmas 9.1.5 and 9.1.6 that

$$H_\mu \left( \mathcal{P}_k / \bigvee_{j=1}^n f^{-j}(\mathcal{P}) \right) \leq H_\mu(\mathcal{P}_k / \mathcal{R}_n) < \varepsilon / 2$$

for every  $n \geq n_0$ . Using Exercise 9.1.1, we get that  $H_\mu(\mathcal{P} / \bigvee_{j=1}^n f^{-j}(\mathcal{P})) < \varepsilon$  for every  $n \geq n_0$ . In this way, it is shown that  $H_\mu(\mathcal{P} / \bigvee_{j=1}^n f^{-j}(\mathcal{P})) \rightarrow 0$ . By Lemma 9.1.12, it follows that  $h_\mu(f, \mathcal{P}) = 0$ .

As a consequence, we get that every system with entropy zero is invertible at almost every point:

**Proposition 9.5.5.** *Let  $(M, \mathcal{B}, \mu)$  be a Lebesgue space and  $f : M \rightarrow M$  be a measure-preserving transformation. If  $h_\mu(f) = 0$  then  $(f, \mu)$  is invertible: there exists a measurable transformation  $g : M \rightarrow M$  that preserves the measure  $\mu$  and satisfies  $f \circ g = g \circ f = \text{id}$  at  $\mu$ -almost every point.*

*Proof.* Consider the homomorphism  $\tilde{f} : \tilde{\mathcal{B}} \rightarrow \tilde{\mathcal{B}}$  induced by  $f$  in the measure algebra of  $\mathcal{B}$  (these notions were introduced in Section 8.5). Recall that  $\tilde{f}$  is always injective (Exercise 8.5.2). Given any  $B \in \mathcal{B}$ , consider the partition  $\mathcal{P} = \{B, B^c\}$ . The hypothesis  $h_\mu(f) = 0$  implies that  $h_\mu(f, \mathcal{P}) = 0$ . By Lemma 9.5.4, it follows that  $\mathcal{P} \prec \bigvee_{j=1}^\infty f^{-j}(\mathcal{P})$ . This implies that  $\mathcal{P} \subset f^{-1}(\mathcal{B})$ , because  $f^{-j}(\mathcal{P}) \subset f^{-1}(\mathcal{B})$  for every  $j \geq 1$ . Varying  $B$ , we conclude that  $\mathcal{B} \subset f^{-1}(\mathcal{B})$ . In other words, the homomorphism  $\tilde{f}$  is surjective. Hence,  $\tilde{f}$  is an isomorphism of measure algebras. Then, by Proposition 8.5.6, there exists a measurable map  $g : M \rightarrow M$  preserving the measure  $\mu$  and such that the corresponding homomorphism of measure algebra  $\tilde{g} : \tilde{\mathcal{B}} \rightarrow \tilde{\mathcal{B}}$  is the inverse of  $\tilde{f}$ . In other words,  $\tilde{f} \circ \tilde{g} = \tilde{g} \circ \tilde{f} = \text{id}$ . Then, (Exercise 8.5.2)  $f \circ g = g \circ f = \text{id}$ , as claimed.

These arguments also prove the following fact, which will be useful in a while:

**Corollary 9.5.6.** *In the conditions of Proposition 9.5.5, every  $\sigma$ -algebra  $\mathcal{A} \subset \mathcal{B}$  that satisfies  $f^{-1}(\mathcal{A}) \subset \mathcal{A}$  up to measure zero also satisfies  $f^{-1}(\mathcal{A}) = \mathcal{A}$  up to measure zero.*

Exercise 9.1.5 implies that if  $(f, \mu)$  has entropy zero then the same is true for any factor. Therefore, the following fact is also an immediate consequence of the proposition:

**Corollary 9.5.7.** *In the conditions of Proposition 9.5.5, every factor of  $(f, \mu)$  is invertible at almost every point.*

It is not completely understood how entropy relates to the spectrum type of a system, but there are several partial results, especially for systems with entropy zero.

Rokhlin [Rok67a, § 14] proved that every ergodic system with discrete spectrum defined in a Lebesgue space has entropy zero. This may also be deduced from the fact that, as we mentioned in Section 8.3, every ergodic system with discrete spectrum is ergodically isomorphic to a translation in a compact abelian group. As we saw in Corollary 8.5.7, in Lebesgue spaces ergodic isomorphism implies ergodic equivalence. Recall also that systems with discrete spectrum in Lebesgue spaces are always invertible (Exercise 8.5.5).

In that same work of Rokhlin it is shown that invertible systems with singular spectrum defined in Lebesgue spaces have entropy zero and the same holds for systems with Lebesgue spectrum of finite rank (if they exist). The case of infinite rank is the focus of the next section. We are going to see that there are systems with Lebesgue spectrum of infinite rank and entropy zero. On the other hand, we introduce the important class of so-called Kolmogorov systems, for which the entropy is necessarily positive, in a strong sense.

### 9.5.3 Kolmogorov systems

Let  $(M, \mathcal{B}, \mu)$  be a non-trivial probability space, that is, one such that not all measurable sets have measure 0 or 1. We use  $\bigvee_{\alpha} \mathcal{U}_{\alpha}$  to denote the  $\sigma$ -algebra generated by any family of subsets  $\mathcal{U}_{\alpha}$  of  $\mathcal{B}$ . Let  $f : M \rightarrow M$  be a transformation preserving the measure  $\mu$ .

**Definition 9.5.8.** We say that  $(f, \mu)$  is a *Kolmogorov system* if there exists some  $\sigma$ -algebra  $\mathcal{A} \subset \mathcal{B}$  such that

- (i)  $f^{-1}(\mathcal{A}) \subset \mathcal{A}$  up to measure zero;
- (ii)  $\bigcap_{n=0}^{\infty} f^{-n}(\mathcal{A}) = \{\emptyset, M\}$  up to measure zero;
- (iii)  $\bigvee_{n=0}^{\infty} \{B \in \mathcal{B} : f^{-n}(B) \in \mathcal{A}\} = \mathcal{B}$  up to measure zero.

We leave it to the reader to check that this property is an invariant of ergodic equivalence (it is *not* an invariant of spectral equivalence, as will be explained shortly).

If  $(f, \mu)$  is a Kolmogorov system then  $(f^k, \mu)$  is a Kolmogorov system, for every  $k \geq 1$ . Indeed, if  $\mathcal{A}$  satisfies condition (i) then the sequence  $f^{-j}(\mathcal{A})$  is non-increasing and, in particular,  $f^{-k}(\mathcal{A}) \subset \mathcal{A}$ . Then, the conditions (ii) and (iii) imply that

$$\bigcap_{n=0}^{\infty} f^{-kn}(\mathcal{A}) = \bigcap_{n=0}^{\infty} f^{-n}(\mathcal{A}) = \{\emptyset, M\} \quad \text{and}$$

$$\bigvee_{n=0}^{\infty} \{B \in \mathcal{B} : f^{-kn}(B) \in \mathcal{A}\} = \bigvee_{n=0}^{\infty} \{B \in \mathcal{B} : f^{-n}(B) \in \mathcal{A}\} = \mathcal{B}$$

up to measure zero. We say that  $(f, \mu)$  is a *Kolmogorov automorphism* if it is invertible and a Kolmogorov system. Then the inverse  $(f^{-1}, \mu)$  is also a Kolmogorov system, as we will see.

**Proposition 9.5.9.** *Every Kolmogorov system has Lebesgue spectrum of infinite rank. If the  $\sigma$ -algebra  $\mathcal{B}$  is countably generated then the rank is countable.*

*Proof.* Let  $\mathcal{A} \subset \mathcal{B}$  be a  $\sigma$ -algebra satisfying the conditions in Definition 9.5.8. Let  $E = L_0^2(M, \mathcal{A}, \mu)$  be the subspace of functions  $\varphi \in L_0^2(M, \mathcal{B}, \mu)$  that are  $\mathcal{A}$ -measurable, that is, such that the pre-image  $\varphi^{-1}(B)$  of every measurable set  $B \subset \mathbb{R}$  is in  $\mathcal{A}$  up to measure zero. We are going to show that  $E$  satisfies all the conditions in Definition 8.4.1.

Start by observing that  $U_f(L_0^2(M, \mathcal{A}, \mu)) = L_0^2(M, f^{-1}(\mathcal{A}), \mu)$ . Indeed, it is clear that if  $\varphi$  is  $\mathcal{A}$ -measurable then  $U_f\varphi = \varphi \circ f$  is  $f^{-1}(\mathcal{A})$ -measurable. The inclusion  $\subset$  follows immediately. Conversely, given any  $B \in f^{-1}(\mathcal{A})$ , take  $A \in \mathcal{A}$  such that  $B = f^{-1}(A)$  and let  $c = \mu(A) = \mu(B)$ . Then,  $\chi_B - c = U_f(\chi_A - c)$  is in  $U_f(L_0^2(M, \mathcal{A}, \mu))$ . This gives the other inclusion. So, the hypothesis that  $f^{-1}(\mathcal{A}) \subset \mathcal{A}$  up to measure zero ensures that  $U_f(E) \subset E$ .

It also follows that  $U_f^n(L_0^2(M, \mathcal{A}, \mu)) = L_0^2(M, f^{-n}(\mathcal{A}), \mu)$  for every  $n \geq 0$ . Then,

$$\bigcap_{n=0}^{\infty} U_f^n(L_0^2(M, \mathcal{A}, \mu)) = L_0^2\left(M, \bigcap_{n=0}^{\infty} f^{-n}(\mathcal{A}), \mu\right).$$

Hence, the hypothesis that  $\bigcap_{n=0}^{\infty} f^{-n}(\mathcal{A}) = \{0, M\}$  up to measure zero implies that  $\bigcap_{n=0}^{\infty} U_f^n(E) = \{0\}$ .

Now consider  $\mathcal{A}_n = \{B \in \mathcal{B} : f^{-n}(B) \in \mathcal{A}\}$ . The sequence  $(\mathcal{A}_n)_n$  is non-decreasing, because  $f^{-1}(\mathcal{A}) \subset \mathcal{A}$ . Moreover, each  $\varphi$  is  $\mathcal{A}_n$ -measurable if and only if  $U_f^n\varphi = \varphi \circ f^n$  is  $\mathcal{A}$ -measurable. This shows that  $U_f^{-n}(L_0^2(M, \mathcal{A}, \mu)) = L_0^2(M, \mathcal{A}_n, \mu)$  for every  $n \geq 0$ . Observe also that

$$\sum_{n=0}^{\infty} L_0^2(M, \mathcal{A}_n, \mu) = L_0^2\left(M, \bigvee_{n=0}^{\infty} \mathcal{A}_n, \mu\right). \tag{9.5.4}$$

Indeed, it is clear that  $L_0^2(M, \mathcal{A}_k, \mu) \subset L_0^2(M, \bigvee_{n=0}^{\infty} \mathcal{A}_n, \mu)$  for every  $k$ , since  $\mathcal{A}_k$  is contained in  $\bigvee_{n=0}^{\infty} \mathcal{A}_n$ . The inclusion  $\subset$  in (9.5.4) is an immediate consequence of this observation, since  $L_0^2(M, \bigvee_{n=0}^{\infty} \mathcal{A}_n, \mu)$  is a Banach space. Now consider any  $A \in \bigvee_{n=0}^{\infty} \mathcal{A}_n$ . The approximation theorem (Theorem A.1.19) implies that for each  $\varepsilon > 0$  there exist  $n$  and  $A_n \in \mathcal{A}_n$  such that  $\mu(A \Delta A_n) < \varepsilon$ . Then  $(\chi_{A_n})_n$  converges to  $\chi_A$  in the  $L^2$ -norm, and so  $\chi_A \in \sum_{n=0}^{\infty} L_0^2(M, \mathcal{A}_n, \mu)$ .

This proves the inclusion  $\supset$  in (9.5.4). In this way, we have shown that

$$\sum_{n=0}^{\infty} U_f^{-n}(L_0^2(M, \mathcal{A}, \mu)) = L_0^2\left(M, \bigvee_{n=0}^{\infty} \mathcal{A}_n, \mu\right).$$

Therefore, the hypothesis that  $\bigvee_{n=0}^{\infty} \mathcal{A}_n = \mathcal{B}$  up to measure zero implies that  $\sum_{n=0}^{\infty} U_f^{-n}(E) = L_0^2(M, \mathcal{B}, \mu)$ .

This concludes the proof that  $(f, \mu)$  has Lebesgue spectrum. To prove that the rank is infinite, we need the following lemma:

**Lemma 9.5.10.** *Let  $\mathcal{A}$  be any  $\sigma$ -algebra satisfying the conditions in Definition 9.5.8. Then for every  $A \in \mathcal{A}$  with  $\mu(A) > 0$  there exists  $B \subset A$  such that  $0 < \mu(B) < \mu(A)$ .*

*Proof.* Suppose that  $\mathcal{A}$  has any element  $A$  with positive measure that does not satisfy the conclusion of the lemma. We claim that  $A \cap f^{-k}(A)$  has measure zero for every  $k \geq 1$ . Then,

$$\mu\left(f^{-i}(A) \cap f^{-j}(A)\right) = \mu\left(A \cap f^{-j+i}(A)\right) = 0 \quad \text{for every } 0 \leq i < j.$$

Since  $\mu(f^{-j}(A)) = \mu(A)$  for every  $j \geq 0$ , this implies that the measure  $\mu$  is infinite, which is a contradiction. This contradiction reduces the proof of the lemma to proving our claim.

To do that, note that condition (i) implies that  $f^{-k}(A) \in f^{-k}(\mathcal{A}) \subset \mathcal{A}$ . Then it follows from the choice of  $A$  that  $A \cap f^{-k}(A)$  must have either zero measure or full measure in  $A$ :

$$\mu\left(A \cap f^{-k}(A)\right) = 0 \quad \text{or else} \quad \mu\left(A \setminus f^{-k}(A)\right) = 0.$$

So, to prove the claim it suffices to exclude the second possibility. Suppose that  $\mu\left(A \setminus f^{-k}(A)\right) = 0$ . Then (Exercise 1.1.4), there exists  $B \in \mathcal{A}$  such that  $\mu(A \Delta B) = 0$  and  $f^{-k}(B) = B$ . It follows that  $B = f^{-nk}(B)$  for every  $n \geq 1$  and, thus,

$$B \in \bigcap_{n \in \mathbb{N}} f^{-nk}(\mathcal{A}) = \bigcap_{n \in \mathbb{N}} f^{-n}(\mathcal{A}).$$

By condition (ii), this means that the measure of  $B$  is either 0 or 1. Since  $\mu(B) = \mu(A)$  is positive, it follows that  $\mu(A) = \mu(B) = 1$ . Then, the hypothesis about  $A$  implies that the  $\sigma$ -algebra  $\mathcal{A}$  contains only sets with measure 0 or 1. By condition (iii), it follows that the same is true for the  $\sigma$ -algebra  $\mathcal{B}$ , which contradicts the assumption that the probability space is non-trivial.

On the way toward proving that the orthogonal complement  $F = E \ominus U_f(E)$  has infinite dimension, let us start by checking that  $F \neq \{0\}$ . Indeed, otherwise we would have  $U_f(E) = E$  and, thus,  $U_f^n(E) = E$  for every  $n \geq 1$ . By condition (ii), that would imply that  $E = \bigcap_n U_f^n(E) = \{0\}$ . Then, by condition (iii), we would have  $L_0^2(M, \mathcal{B}, \mu) = \{0\}$  and that would contradict the hypothesis that the probability space is non-trivial.



Let  $\varphi$  be any non-zero element of  $F$ , fixed once and for all, and let  $N$  be the set of all  $x \in M$  such that  $\varphi(x) \neq 0$ . Then,  $N \in \mathcal{A}$  and  $\mu(N) > 0$ . It is convenient to consider the space  $E' = L^2(M, \mathcal{A}, \mu) = E \oplus \{\text{constants}\}$ . Observe that  $F$  coincides with  $E' \ominus U_f(E')$ , because the Koopman operator preserves the line of constant functions. Let  $E'_N$  be the subspace of functions  $\psi \in E'$  that vanish outside  $N$ , that is, such that  $\psi(x) = 0$  for every  $x \in N^c$ . By Lemma 9.5.10, we may find sets  $A_j \in \mathcal{A}$ ,  $j \geq 1$  with positive measure, contained in  $N$  and pairwise disjoint. Then,  $\mathcal{X}_{A_j}$  is in  $E'_N$  for every  $j$ . Moreover,  $A_i \cap A_j = \emptyset$  yields  $\mathcal{X}_{A_i} \cdot \mathcal{X}_{A_j} = 0$  for every  $i \neq j$ . This implies that  $E'_N$  has infinite dimension.

Now denote by  $U_f(E')_N$  the subspace of functions  $\psi \in U_f(E')$  that vanish outside  $N$ . Denote  $F_N = E'_N \ominus U_f(E')_N$ . The fact that  $\dim E'_N = \infty$  ensures that  $\dim F_N = \infty$  or  $\dim U_f(E')_N = \infty$  (or both). We are going to show that any of these alternatives implies that  $\dim F = \infty$ .

To treat the first alternative, it suffices to show that  $F_N \subset F$ . For that, since it is clear that  $F_N \subset E'$ , it suffices to check that  $F_N$  is orthogonal to  $U_f(E')$ . Consider any  $\zeta \in F_N$  and  $\eta \in E'$ . The function  $(U_f \eta) \mathcal{X}_N = U_f(\eta \mathcal{X}_{f^{-1}(N)})$  is in  $U_f(E')$  and vanishes outside  $N$ . In other words, it is in  $U_f(E')_N$ . Then  $\zeta \cdot U_f \eta = \zeta \cdot (U_f \eta) \mathcal{X}_N = 0$ , because  $\zeta$  vanishes outside  $N$  and is orthogonal to  $U_f(E')_N$ . This completes the argument in this case.

Now we treat the second alternative. Given any  $U_f \eta \in U_f(E')_N$  and any  $n \in \mathbb{N}$ , let  $\eta_n = \eta \mathcal{X}_{R_n}$  with  $R_n = \{x \in M : |\eta(x)| \leq n\}$ . Then  $(\eta_n)_n$  is a sequence of bounded functions converging to  $\eta$  in  $E'$ . Moreover, every  $\eta_n$  vanishes outside  $f^{-1}(N)$ , because  $\eta$  does. Then,  $(U_f \eta_n)_n$  is a sequence of bounded functions that vanish outside  $N$  and, recalling that  $U_f$  is an isometry, this sequence converges to  $U_f \eta$  in  $E'$ . This proves that the subspace of bounded functions is dense in  $U_f(E')_N$ . Then, since we are assuming that  $\dim U_f(E')_N = \infty$ , this subspace must also have infinite dimension. Choose  $\{\zeta_k : k \geq 1\} \subset E'$  such that  $\{U_f \zeta_k : k \geq 1\}$  is a linearly independent subset of  $U_f(E')_N$  consisting of bounded functions. Observe that the products  $\varphi(U_f \zeta_k)$ ,  $k \geq 1$  form a linearly independent subset of  $E'$ . Moreover, given any  $\eta \in E'$ ,

$$\varphi(U_f \zeta_k) \cdot (U_f \eta) = \int \varphi(\zeta_k \circ f) (\bar{\eta} \circ f) d\mu = \int \varphi(\zeta_k \bar{\eta}) \circ f d\mu = \varphi \cdot U_f(\bar{\zeta}_k \eta).$$

This last expression is equal to zero because  $\bar{\zeta}_k \eta \in E'$  and the function  $\varphi \in F$  is orthogonal to  $U_f(E')$ . Varying  $\eta \in E'$ , we conclude that  $\varphi(U_f \zeta_k)$  is orthogonal to  $U_f(E')$  for every  $k$ . This shows that  $\{\varphi(U_f \zeta_k) : k \geq 1\}$  is contained in  $F$  and, thus,  $\dim F = \infty$  also in this case.

This completes the proof that  $(f, \mu)$  has infinite rank. When  $\mathcal{B}$  is countably generated,  $L^2_0(M, \mathcal{B}, \mu)$  is separable (Example 8.4.7) and so the rank is necessarily countable.

We say that a partition of  $(M, \mathcal{B}, \mu)$  is trivial if all its elements have measure 0 or 1. Keep in mind that in the present chapter all partitions are assumed to be countable.

**Proposition 9.5.11.** *A system  $(f, \mu)$  in a Lebesgue space is a Kolmogorov system if and only if  $h_\mu(f, \mathcal{P}) > 0$  for every non-trivial partition with finite entropy. In particular, every Kolmogorov system has positive entropy.*

This result is due to Pinsker [Pin60] and to Rokhlin and Sinai [RS61]. The proof may also be found in Rokhlin [Rok67a, §13]. Let us point out, however, that the last part of the statement is an immediate consequence of the ideas in Section 9.5.2. Indeed, suppose that  $(f, \mu)$  is a Kolmogorov system with zero entropy. By Corollary 9.5.6, any  $\sigma$ -algebra  $\mathcal{A}$  that satisfies condition (i) in Definition 9.5.8 also satisfies  $f^{-1}(\mathcal{A}) = \mathcal{A}$  up to measure zero. Then, condition (ii) implies that  $\mathcal{A}$  is trivial and, by condition (iii), the  $\sigma$ -algebra  $\mathcal{B}$  itself is trivial (contradicting the assumption we made at the beginning of this section).

It follows from Proposition 9.5.11 and the relation (9.1.21) that the inverse of a Kolmogorov automorphism is also a Kolmogorov automorphism. Unlike what happens for Bernoulli automorphisms (Exercise 9.5.1), in the Kolmogorov case the two systems  $(f, \mu)$  and  $(f^{-1}, \mu)$  need not be ergodically equivalent.

**Example 9.5.12.** The first example of an invertible system with countable Lebesgue spectrum that is not a Kolmogorov system was found by Girsanov in 1959, but was never published. Another example, a factor of a certain Gaussian shift (recall Example 8.4.13) with countable Lebesgue spectrum but whose entropy vanishes, was exhibited a few years later by Newton and Parry [NP66]. Also, Gurevič [Gur61] proved that the horocyclic flow on surfaces with constant negative curvature has entropy zero; sometime before, Parasyuk [Par53] had shown that such flows have countable Lebesgue spectrum.

As we saw in Theorem 8.4.11, all systems with countable Lebesgue spectrum are spectrally equivalent. Therefore, the fact that systems as in Example 9.5.12 do exist has the interesting consequence that being a Kolmogorov system is *not* an invariant of spectral equivalence.

**Example 9.5.13.** We saw in Examples 8.4.2 and 8.4.3 that all the Bernoulli shifts have Lebesgue spectrum. In both cases, one-sided and two-sided, we exhibited subspaces of  $L_0^2(M, \mathcal{B}, \mu)$  of the form  $E = L_0^2(M, \mathcal{A}, \mu)$  for some  $\sigma$ -algebra  $\mathcal{A} \subset \mathcal{B}$ . Therefore, the same argument proves that every Bernoulli shift is a Kolmogorov system. In particular, every Bernoulli automorphism is a Kolmogorov automorphism.

There exist Kolmogorov automorphisms that are not Bernoulli automorphisms. The first example, found by Ornstein, is quite elaborate. The following simple construction is due to Kalikow [Kal82]:

**Example 9.5.14.** Let  $\sigma : \Sigma \rightarrow \Sigma$  be the shift map in  $\Sigma = \{1, 2\}^{\mathbb{Z}}$  and  $\mu$  be the Bernoulli measure associated with the probability vector  $p = (1/2, 1/2)$ .

Consider the map  $f: \Sigma \times \Sigma \rightarrow \Sigma \times \Sigma$  defined as follows:

$$f((x_n)_n, (y_n)_n) = (\sigma^{\pm 1}((x_n)_n), \sigma^{\pm 1}((y_n)_n))$$

where the sign is  $-$  if  $x_0 = 1$  and  $+$  if  $x_0 = 2$ . This map preserves the product measure  $\mu \times \mu$ . Kalikow showed that  $(f, \mu)$  is a Kolmogorov automorphism but not a Bernoulli automorphism.

Consider any Kolmogorov automorphism that is not a Bernoulli automorphism and let  $s > 0$  be its entropy. Consider any Bernoulli automorphism whose entropy is equal to  $s$  (see Exercise 9.2.2). The two systems have the same entropy but they cannot be ergodically equivalent, since being a Bernoulli automorphism is an invariant of ergodic equivalence. Therefore, the entropy is not a complete invariant of ergodic equivalence for Kolmogorov automorphisms. Actually (see Ornstein and Shields [OS73]), there exists a non-countable family of Kolmogorov automorphisms that have the same entropy and, yet, are not ergodically equivalent.

The properties of Bernoulli automorphisms described in Exercise 9.5.1 do not extend to the Kolmogorov case: there exist Kolmogorov automorphisms that are not ergodically equivalent to their inverses (see Ornstein and Shields [OS73]), and there are also Kolmogorov automorphisms that admit no  $k$ -th root for any value of  $k \geq 1$  (Clark [Cla72]).

Closing this section, let us discuss the Kolmogorov property for two specific classes of systems: Markov shifts and automorphisms of compact groups.

Concerning the first class, Friedman and Ornstein [FO70] proved that every two-sided mixing Markov shift is a Bernoulli automorphism. Recall (Theorem 7.2.11) that a Markov shift is mixing if and only if the corresponding stochastic matrix is aperiodic. It follows from the theorem of Friedman and Ornstein that the entropy is still a complete invariant of ergodic equivalence in the context of two-sided mixing Markov shifts. Another interesting consequence is that every two-sided mixing Markov shift is a Kolmogorov automorphism. Observe, however, that this consequence admits a relatively easy direct proof (see Exercise 9.5.4).

As for the second class, every ergodic automorphism of a compact group is a Kolmogorov automorphism. This was proven by Rokhlin [Rok67b] for abelian groups and by Yuzvinskii [Yuz68] in the general case. In fact, ergodic automorphisms of metrizable compact groups are Bernoulli automorphisms (Lind [Lin77], Miles and Thomas [MT78]). In particular, every ergodic linear automorphism of the torus  $\mathbb{T}^d$  is a Bernoulli automorphism; this had been proved by Katznelson [Kat71]. Recall (Theorem 4.2.14) that a linear automorphism  $f_A$  is ergodic if and only if no eigenvalue of the matrix  $A$  is a root of unity.

### 9.5.4 Exact systems

We say that a Kolmogorov system is *exact* if one may take the  $\sigma$ -algebra  $\mathcal{A}$  in Definition 9.5.8 to be the  $\sigma$ -algebra  $\mathcal{B}$  of all measurable sets. Note that in this case the conditions (i) and (iii) are automatically satisfied. Therefore, a system  $(f, \mu)$  is exact if and only if the  $\sigma$ -algebra  $\mathcal{B}$  is such that  $\bigcap_{n=0}^{\infty} f^{-n}(\mathcal{B})$  is trivial, meaning that it only contains sets with measure 0 or 1. Equivalently,  $(f, \mu)$  is exact if and only if

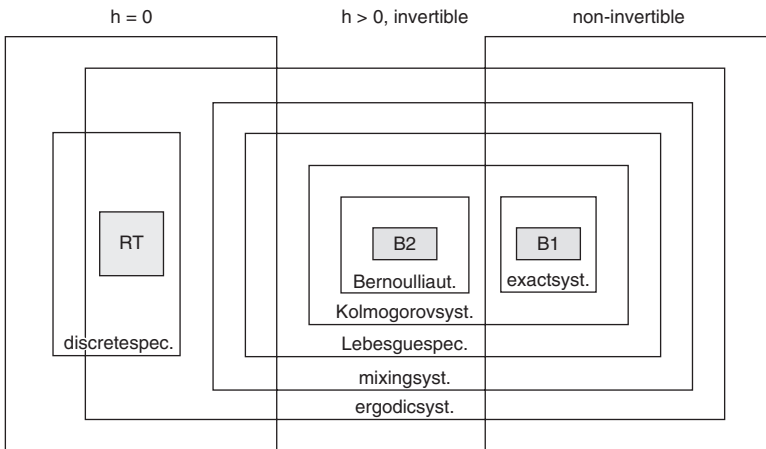
$$\bigcap_{n=0}^{\infty} U_f^n(L_0^2(M, \mathcal{B}, \mu)) = \{0\}.$$

This observation also implies that, unlike the Kolmogorov property, exactness is an invariant of spectral equivalence.

We saw in Example 8.4.2 that every one-sided Bernoulli shift has Lebesgue spectrum. In order to do that, we considered the subspace  $E = L_0^2(M, \mathcal{B}, \mu)$ . Therefore, the same argument proves that every one-sided Bernoulli shift is an exact system. A much larger class of examples, expanding maps endowed with their equilibrium states, is studied in Chapter 12.

It is immediate that invertible systems are never exact: in the invertible case  $f^{-n}(\mathcal{B}) = \mathcal{B}$  up to measure zero, for every  $n$ ; therefore, the exactness condition corresponds to saying that the  $\sigma$ -algebra  $\mathcal{B}$  is trivial (which is excluded, by hypothesis).

Figure 9.2 summarizes the relations between the different classes of systems studied in this book. It is organized in three columns: systems with zero entropy (which are necessarily invertible, as we saw in Proposition 9.5.5), invertible systems with positive entropy and non-invertible systems.



**Figure 9.2.** Relations between various of classes of systems (B1/B2 = one-sided/two-sided Bernoulli shifts, RT = irrational rotations on tori)

### 9.5.5 Exercises

- 9.5.1. Show that if  $(f, \mu)$  is a Bernoulli automorphism then it is ergodically equivalent to its inverse  $(f^{-1}, \mu)$ . Moreover, for every  $k \geq 1$  there exists a Bernoulli automorphism  $(g, \nu)$  that is a  $k$ -th root of  $(f, \mu)$ , that is, such that  $(g^k, \nu)$  is ergodically equivalent to  $(f, \mu)$ . [Observation: Ornstein [Orn74] proved that, conversely, every  $k$ -th root of a Bernoulli automorphism is a Bernoulli automorphism.]
- 9.5.2. Use the notion of density point to show that the decimal expansion map  $f(x) = 10x - [10x]$  is exact, relative to the Lebesgue measure.
- 9.5.3. Show that the Gauss map is exact, relative to its absolutely continuous invariant measure  $\mu$ .
- 9.5.4. Show that the two-sided Markov shift associated with any aperiodic stochastic matrix  $P$  is a Kolmogorov automorphism.
- 9.5.5. Show that the one-sided Markov shift associated with any aperiodic stochastic matrix  $P$  is an exact system.
- 9.5.6. Prove that if  $(f, \mu)$  is exact then  $h_\mu(f, \mathcal{P}) > 0$  for every non-trivial partition  $\mathcal{P}$  with finite entropy.

## 9.6 Entropy and ergodic decomposition

It is not difficult to show that the entropy  $h_\mu(f)$  is always an *affine* function of the invariant measure  $\mu$ :

**Proposition 9.6.1.** *Let  $\mu$  and  $\nu$  be probability measures invariant under a transformation  $f : M \rightarrow M$ . Then,  $h_{t\mu+(1-t)\nu}(f) = th_\mu(f) + (1-t)h_\nu(f)$  for every  $0 < t < 1$ .*

*Proof.* Define  $\phi(x) = -x \log x$  for  $x > 0$ . On the one hand, since the function  $\phi$  is concave,

$$\phi(t\mu(B) + (1-t)\nu(B)) \geq t\phi(\mu(B)) + (1-t)\phi(\nu(B))$$

for every measurable set  $B \subset M$ . On the other hand, given any measurable set  $B \subset M$ ,

$$\begin{aligned} & \phi(t\mu(B) + (1-t)\nu(B)) - t\phi(\mu(B)) - (1-t)\phi(\nu(B)) \\ &= -t\mu(B) \log \frac{t\mu(B) + (1-t)\nu(B)}{\mu(B)} - (1-t)\nu(B) \log \frac{t\mu(B) + (1-t)\nu(B)}{\nu(B)} \\ &\leq -t\mu(B) \log t - (1-t)\nu(B) \log(1-t) \end{aligned}$$

because the function  $-\log$  is decreasing. Therefore, given any partition  $\mathcal{P}$  with finite entropy,

$$H_{t\mu+(1-t)\nu}(\mathcal{P}) \geq tH_\mu(\mathcal{P}) + (1-t)H_\nu(\mathcal{P}) \quad \text{and}$$

$$H_{t\mu+(1-t)\nu}(\mathcal{P}) \leq tH_\mu(\mathcal{P}) + (1-t)H_\nu(\mathcal{P}) - t \log t - (1-t) \log(1-t).$$

Consequently,

$$h_{t\mu+(1-t)\nu}(f, \mathcal{P}) = th_{\mu}(f, \mathcal{P}) + (1-t)h_{\nu}(f, \mathcal{P}). \tag{9.6.1}$$

It follows, immediately, that  $h_{t\mu+(1-t)\nu}(f) \leq th_{\mu}(f) + (1-t)h_{\nu}(f)$ . Moreover, the relations (9.1.16) and (9.6.1) imply that

$$h_{t\mu+(1-t)\nu}(f, \mathcal{P}_1 \vee \mathcal{P}_2) \geq th_{\mu}(f, \mathcal{P}_1) + (1-t)h_{\nu}(f, \mathcal{P}_2)$$

for any partitions  $\mathcal{P}_1$  and  $\mathcal{P}_2$ . Taking the supremum on  $\mathcal{P}_1$  and  $\mathcal{P}_2$  we obtain that  $h_{t\mu+(1-t)\nu}(f) \geq th_{\mu}(f) + (1-t)h_{\nu}(f)$ .

In particular, given any invariant set  $A \subset M$ , we have that

$$h_{\mu}(f) = \mu(A)h_{\mu_A}(f) + \mu(A^c)h_{\mu_{A^c}}(f), \tag{9.6.2}$$

where  $\mu_A$  and  $\mu_{A^c}$  denote the normalized restrictions of  $\mu$  to the set  $A$  and its complement, respectively (this fact was obtained before, in Exercise 9.1.4). Another immediate consequence is the following version of Proposition 9.6.1 for finite convex combinations:

$$\mu = \sum_{i=1}^n t_i \mu_i \quad \Rightarrow \quad h_{\mu}(f) = \sum_{i=1}^n t_i h_{\mu_i}(f), \tag{9.6.3}$$

for any invariant probability measures  $\mu_1, \dots, \mu_n$  and any positive numbers  $t_1, \dots, t_n$  with  $\sum_{i=1}^n t_i = 1$ .

A much deeper fact, due to Konrad Jacobs [Jac60, Jac63], is that the affinity property extends to the ergodic decomposition given by Theorem 5.1.3:

**Theorem 9.6.2** (Jacobs). *Suppose that  $M$  is a complete separable metric space. Given any invariant probability measure  $\mu$ , let  $\{\mu_P : P \in \mathcal{P}\}$  be its ergodic decomposition. Then,  $h_{\mu}(f) = \int h_{\mu_P}(f) d\hat{\mu}(P)$  (if one side is infinite, so is the other side).*

We are going to deduce this result from a general theorem about affine functionals in the space of probability measures, that we state in Section 9.6.1 and prove in Section 9.6.2.

### 9.6.1 Affine property

Let  $M$  be a complete separable metric space. We saw in Lemma 2.1.3 that the weak\* topology in the space of probability measures  $\mathcal{M}_1(M)$  is metrizable. Moreover (Exercise 2.1.3), the metric space  $\mathcal{M}_1(M)$  is separable.

Let  $W$  be a probability measure on the Borel  $\sigma$ -algebra of  $\mathcal{M}_1(M)$ . The *barycenter* of  $W$  is the probability measure  $\text{bar}(W) \in \mathcal{M}_1(M)$  given by

$$\int \psi d\text{bar}(W) = \int \left( \int \psi d\eta \right) dW(\eta) \tag{9.6.4}$$

for every bounded measurable function  $\psi : M \rightarrow \mathbb{R}$ . We leave it to the reader to check that this relation determines  $\text{bar}(W)$  uniquely (Exercise 9.6.1) and that the barycenter is an affine function of the measure (Exercise 9.6.2).

**Example 9.6.3.** If  $W$  is a Dirac measure, that is, if  $W = \delta_\nu$  for some  $\nu \in \mathcal{M}_1(M)$ , then  $\text{bar}(W) = \nu$ . Using Exercise 9.6.2, we get the following generalization: if  $W = \sum_{i=1}^\infty t_i \delta_{\nu_i}$  with  $t_i \geq 0$  and  $\sum_{i=1}^\infty t_i = 1$  and  $\nu_i \in \mathcal{M}_1(M)$  for every  $i$ , then  $\text{bar}(W) = \sum_{i=1}^\infty t_i \nu_i$ .

**Example 9.6.4.** Let  $\{\mu_P : P \in \mathcal{P}\}$  be the ergodic decomposition of a probability measure  $\mu$  invariant under a measurable transformation  $f : M \rightarrow M$  and  $\hat{\mu}$  be the associated quotient measure in  $\mathcal{P}$  (recall Section 5.1.1). Let  $W = \Xi_* \hat{\mu}$  be the image of the quotient measure  $\hat{\mu}$  under the map  $\Xi : \mathcal{P} \rightarrow \mathcal{M}$  that assigns to each  $P \in \mathcal{P}$  the conditional probability  $\mu_P$ . Then (Exercise 5.1.4),

$$\int \psi d\mu = \int \left( \int \psi d\mu_P \right) d\hat{\mu}(P) = \int \left( \int \psi d\eta \right) dW(\eta)$$

for every bounded measurable function  $\psi : M \rightarrow \mathbb{R}$ . This means that  $\mu$  is the barycenter of  $W$ .

A set  $\mathcal{M} \subset \mathcal{M}_1(M)$  is said to be *strongly convex* if  $\sum_{i=1}^\infty t_i \nu_i \in \mathcal{M}$  for any  $\nu_i \in \mathcal{M}$  and  $t_i \geq 0$  with  $\sum_{i=1}^\infty t_i = 1$ .

**Theorem 9.6.5.** Let  $\mathcal{M}$  be a strongly convex subset of  $\mathcal{M}_1(M)$  and  $H : \mathcal{M} \rightarrow \mathbb{R}$  be a non-negative affine functional. If  $H$  is upper semi-continuous then

$$H(\text{bar}(W)) = \int H(\eta) dW(\eta)$$

for any probability measure  $W$  on  $\mathcal{M}_1(M)$  with  $W(\mathcal{M}) = 1$  and  $\text{bar}(W) \in \mathcal{M}$ .

Before proving this result, let us explain how Theorem 9.6.2 may be obtained from it. The essential step is the following lemma:

**Lemma 9.6.6.**  $h_\mu(f, \mathcal{Q}) = \int h_{\mu_P}(f, \mathcal{Q}) d\hat{\mu}(P)$  for any finite partition  $\mathcal{Q}$  of  $M$ .

*Proof.* Let  $\mathcal{M} = \mathcal{M}_1(f)$ , the subspace of invariant probability measures, and  $H : \mathcal{M} \rightarrow \mathbb{R}$  be the functional defined by  $H(\eta) = h_\eta(f, \mathcal{Q})$ . Let  $W$  be the image of the quotient measure  $\hat{\mu}$  by the map  $\Xi : \mathcal{P} \rightarrow \mathcal{M}$  that assigns to each  $P \in \mathcal{P}$  the conditional probability  $\mu_P$ . It is clear that  $\mathcal{M}$  is strongly convex,  $W(\mathcal{M}) = 1$  and (recall Example 9.6.4) the barycenter  $\text{bar}(W) = \mu$  is in  $\mathcal{M}$ . Proposition 9.6.1 shows that  $H$  is affine and it is clear that  $H$  is non-negative. In order to apply Theorem 9.6.5, we also need to check that  $H$  is upper semi-continuous.

Initially, suppose that  $f$  is the shift map in a space  $\Sigma = X^\mathbb{N}$ , where  $X$  is a finite set, and  $\mathcal{Q}$  is the partition of  $\Sigma$  into cylinders  $[0; a]$ ,  $a \in X$ . The point with this partition is that its elements are both open and closed subsets of  $\Sigma$ . In other words,  $\partial Q = \emptyset$  for every  $Q \in \mathcal{Q}$ . By Proposition 9.2.12, it follows that

the map  $\eta \mapsto H(\eta) = h_\eta(f, \mathcal{Q})$  is upper semi-continuous at every point of  $\mathcal{M}$ . So, we may indeed apply Theorem 9.6.5 to the functional  $H$ . In this way we get that

$$\begin{aligned} h_\mu(f, \mathcal{Q}) &= H(\mu) = H(\text{bar}(W)) = \int H(\eta) dW(\eta) \\ &= \int H(\mu_P) d\hat{\mu}(P) = \int h_{\mu_P}(f, \mathcal{Q}) d\hat{\mu}(P). \end{aligned}$$

Now we treat the general case of the lemma, by reduction to the previous paragraph. Given any finite partition  $\mathcal{Q}$ , let  $\Sigma = \mathcal{Q}^{\mathbb{N}}$  and

$$h : M \rightarrow \Sigma, \quad h(x) = (\mathcal{Q}(f^n(x)))_{n \in \mathbb{N}}.$$

Observe that  $h \circ f = \sigma \circ h$ , where  $\sigma : \Sigma \rightarrow \Sigma$  denotes the shift map. To each measure  $\eta$  on  $M$  we may assign the measure  $\eta' = h_*\eta$  on  $\Sigma$ . The previous relation ensures that if  $\eta$  is invariant under  $f$  then  $\eta'$  is invariant under  $\sigma$ . Moreover, if  $\eta$  is ergodic for  $f$  then  $\eta'$  is ergodic for  $\sigma$ . Indeed, if  $B' \subset \Sigma$  is invariant under  $\sigma$  then  $B = h^{-1}(B')$  is invariant under  $f$ . Assuming that  $\eta$  is ergodic, it follows that  $\eta'(B') = \eta(B)$  is either 0 or 1; hence,  $\eta'$  is ergodic.

By construction,  $\mathcal{Q} = h^{-1}(\mathcal{Q}')$ , where  $\mathcal{Q}'$  denotes the partition of  $\Sigma$  into the cylinders  $[0; \mathcal{Q}]$ ,  $Q \in \mathcal{Q}$ . More generally,  $\bigvee_{j=0}^{n-1} f^{-j}(\mathcal{Q}) = h^{-1}(\bigvee_{j=0}^{n-1} \sigma^{-j}(\mathcal{Q}'))$  and, thus,

$$H_\eta \left( \bigvee_{j=0}^{n-1} f^{-j}(\mathcal{Q}) \right) = H_{\eta'} \left( \bigvee_{j=0}^{n-1} \sigma^{-j}(\mathcal{Q}') \right)$$

for every  $n \in \mathbb{N}$ . Dividing by  $n$  and taking the limit,

$$h_\eta(f, \mathcal{Q}) = h_{\eta'}(\sigma, \mathcal{Q}') \quad \text{for every } \eta \in \mathcal{M}. \tag{9.6.5}$$

Denote  $\mu' = h_*\mu$  and  $\mu'_P = h_*(\mu_P)$  for each  $P$ . For every bounded measurable function  $\psi : \Sigma \rightarrow \mathbb{R}$ ,

$$\begin{aligned} \int \psi d\mu' &= \int (\psi \circ h) d\mu = \int \left( \int (\psi \circ h) d\mu_P \right) d\hat{\mu}(P) \\ &= \int \left( \int \psi d\mu'_P \right) d\hat{\mu}(P). \end{aligned} \tag{9.6.6}$$

As the measures  $\mu'_P$  are ergodic, the relation (9.6.6) means that  $\{\mu'_P : P \in \mathcal{P}\}$  is an ergodic decomposition of  $\mu'$ . Then, according to the previous paragraph,  $h_{\mu'}(\sigma, \mathcal{Q}') = \int h_{\mu'_P}(\sigma, \mathcal{Q}') d\hat{\mu}(P)$ . By the relation (9.6.5) applied to  $\eta = \mu$  and to  $\eta = \mu_P$ , this may be rewritten as

$$h_\mu(\sigma, \mathcal{Q}) = \int h_{\mu_P}(\sigma, \mathcal{Q}) d\hat{\mu}(P),$$

which is precisely what we wanted to prove.

Proceeding with the proof of Theorem 9.6.2, consider any increasing sequence  $\mathcal{Q}_1 \prec \dots \prec \mathcal{Q}_n \prec \dots$  of finite partitions of  $M$  such that  $\text{diam } \mathcal{Q}_n(x)$



converges to zero at every  $x \in M$  (such a sequence may be constructed, for instance, from a family of balls centered at the points of a countable dense subset, with radii converging to zero). By Lemma 9.6.6,

$$h_\mu(f, \mathcal{Q}_n) = \int h_{\mu_P}(f, \mathcal{Q}_n) d\hat{\mu}(P) \tag{9.6.7}$$

for every  $n$ . According to (9.1.16), the sequence  $h_\eta(f, \mathcal{Q}_n)$  is non-decreasing, for any invariant measure  $\eta$ . Moreover, by Corollary 9.2.5, its limit is equal to  $h_\eta(f)$ . Then, we may pass to the limit in (9.6.7) with the aid of the monotone convergence theorem. In this way we get that

$$h_\mu(f) = \int h_{\mu_P}(f) d\hat{\mu}(P),$$

as we wanted to prove. Note that the argument remains valid even when either of the two sides of this identity is infinite (then the other one is also infinite).

In this way, we reduced the proof of Theorem 9.6.2 to proving Theorem 9.6.5.

### 9.6.2 Proof of the Jacobs theorem

Now we prove Theorem 9.6.5. Let us start by proving that the barycenter function has the following continuity property: if  $W$  is concentrated in a neighborhood  $V$  of a given measure  $\nu$  then the barycenter of  $W$  is close to  $\nu$ . More precisely:

**Lemma 9.6.7.** *Let  $W$  be a probability measure on  $\mathcal{M}_1(M)$  and  $\nu \in \mathcal{M}_1(M)$ . Given any finite set  $\Phi = \{\phi_1, \dots, \phi_N\}$  of bounded continuous functions and any  $\varepsilon > 0$ , let  $V = V(\nu, \Phi, \varepsilon)$  be as defined in (2.1.1). If  $W(V) = 1$ , then  $\text{bar}(W) \in V$ .*

*Proof.* Consider any  $i = 1, \dots, N$ . By the definition of barycenter and the hypothesis that the complement of  $V$  has measure zero,

$$\begin{aligned} \left| \int \phi_i d\text{bar}(W) - \int \phi_i d\nu \right| &= \left| \int \left( \int \phi_i d\eta \right) dW(\eta) - \int \left( \int \phi_i d\nu \right) dW(\eta) \right| \\ &\leq \int_V \left| \left( \int \phi_i d\eta - \int \phi_i d\nu \right) \right| dW(\eta). \end{aligned}$$

By the definition of  $V$ , the last expression is smaller than  $\varepsilon$ . Therefore,

$$\left| \int \phi_i d\text{bar}(W) - \int \phi_i d\nu \right| < \varepsilon$$

for every  $i = 1, \dots, N$ . In other words,  $\text{bar}(W) \in V$ .

We also use the following simple property of non-negative affine functionals:

**Lemma 9.6.8.** For any non-negative affine functional  $H : \mathcal{M} \rightarrow \mathbb{R}$ , probability measures  $\nu_i \in \mathcal{M}$ ,  $i \geq 1$  and non-negative numbers  $t_i$ ,  $i \geq 1$  with  $\sum_{i=1}^\infty t_i = 1$ ,

$$H\left(\sum_{i=1}^\infty t_i \nu_i\right) \geq \sum_{i=1}^\infty t_i H(\nu_i).$$

*Proof.* Define  $s_n = \sum_{i=1}^n t_i$  for every  $n \geq 1$ . Let  $R_n = (1 - s_n)^{-1} \sum_{i>n} t_i \nu_i$  if  $s_n < 1$ ; otherwise, pick  $R_n$  arbitrarily. Then,

$$\sum_{i=1}^\infty t_i \nu_i = \sum_{i=1}^n t_i \nu_i + (1 - s_n) R_n.$$

Since  $H$  is affine and the expression on the right-hand side is a (finite) convex combination, it follows that

$$H\left(\sum_{i=1}^\infty t_i \nu_i\right) = \sum_{i=1}^n t_i H(\nu_i) + (1 - s_n) H(R_n) \geq \sum_{i=1}^n t_i H(\nu_i)$$

for every  $n$ . Now just make  $n$  go to infinity.

**Corollary 9.6.9.** If  $H : \mathcal{M} \rightarrow \mathbb{R}$  is a non-negative affine functional then  $H$  is bounded.

*Proof.* Suppose that  $H$  is not bounded: there exist  $\nu_i \in \mathcal{M}$  such that  $H(\nu_i) \geq 2^i$  for every  $i \geq 1$ . Consider  $\nu = \sum_{i=1}^\infty 2^{-i} \nu_i$ . By Lemma 9.6.8,

$$H(\nu) \geq \sum_{i=1}^\infty 2^{-i} H(\nu_i) = \infty.$$

This contradicts the fact that  $H(\nu)$  is finite.

Now we are ready to prove the inequality  $\geq$  in Theorem 9.6.5. Let us write  $\mu = \text{bar}(W)$ . By the hypothesis of semi-continuity, given any  $\varepsilon > 0$  there exist  $\delta > 0$  and a finite family  $\Phi = \{\phi_1, \dots, \phi_N\}$  of bounded continuous functions such that

$$H(\eta) < H(\mu) + \varepsilon \quad \text{for every } \eta \in \mathcal{M} \cap V(\mu, \Phi, \delta). \tag{9.6.8}$$

Since  $\mathcal{M}_1(M)$  is a separable metric space, it admits a countable basis of open sets, and then so does any subspace. Let  $\{V_1, \dots, V_n, \dots\}$  be a basis of open sets of  $\mathcal{M}$ , with the following properties:

- (i) every  $V_n$  is contained in  $\mathcal{M} \cap V(\nu_n, \Phi, \delta)$  for some  $\nu_n \in \mathcal{M}$ ;
- (ii)  $H(\eta) < H(\nu_n) + \varepsilon$  for every  $\eta \in V_n$ .

Consider the partition  $\{P_1, \dots, P_n, \dots\}$  of the space  $\mathcal{M}$  defined by  $P_1 = V_1$  and  $P_n = V_n \setminus (V_1 \cup \dots \cup V_{n-1})$  for every  $n > 1$ . It is clear that the properties (i) and (ii) remain valid if we replace  $V_n$  by  $P_n$ . We claim that

$$\sum_n W(P_n) \nu_n \in V(\mu, \Phi, \delta). \tag{9.6.9}$$

Indeed, observe that

$$\left| \int \phi_i d\mu - \sum_n W(P_n) \int \phi_i dv_n \right| = \left| \sum_n \int_{P_n} \left( \int \phi_i d\eta - \int \phi_i dv_n \right) dW(\eta) \right|$$

for every  $i$ . Therefore, property (i) ensures that

$$\left| \int \phi_i d\mu - \sum_n W(P_n) \int \phi_i dv_n \right| < \sum_n \delta W(P_n) = \delta \quad \text{for every } i,$$

which is precisely what (9.6.9) means. Then, combining (9.6.8), (9.6.9) and Lemma 9.6.8,

$$\sum_n W(P_n)H(v_n) \leq H\left(\sum_n W(P_n)v_n\right) < H(\mu) + \varepsilon.$$

On the other hand, property (ii) implies that

$$\begin{aligned} \int H(\eta) dW(\eta) - \sum_n W(P_n)H(v_n) &= \sum_n \int_{P_n} (H(\eta) - H(v_n)) dW\mu(\eta) \\ &< \sum_n \varepsilon W(P_n) = \varepsilon. \end{aligned}$$

Adding the last two inequalities, we get that  $\int H(\eta) dW(\eta) < H(\mu) + 2\varepsilon$ . Since  $\varepsilon > 0$  is arbitrary, this implies that  $H(\mu) \geq \int H(\eta) dW(\eta)$ .

Now we prove the inequality  $\leq$  in Theorem 9.6.5. Consider any sequence  $(\mathcal{P}_n)_n$  of finite partitions of  $\mathcal{M}$  such that, for every  $v \in \mathcal{M}$ , the diameter of  $\mathcal{P}_n(v)$  converges to zero when  $n$  goes to infinity. For example,  $\mathcal{P}_n = \bigvee_{i=1}^n \{V_i, V_i^c\}$ , where  $\{V_n : n \geq 1\}$  is any countable basis of open sets of  $\mathcal{M}$ . For each fixed  $n$ , consider the normalized restriction  $W_P$  of the measure  $W$  to each set  $P \in \mathcal{P}_n$  (we consider only sets with positive measure: the union of all the elements of  $\bigcup_n \mathcal{P}_n$  with  $W(P) = 0$  has measure zero and so may be neglected):

$$W_P(A) = \frac{W(A \cap P)}{W(P)} \quad \text{for each measurable set } A \subset M.$$

It is clear that  $W = \sum_{P \in \mathcal{P}_n} W(P)W_P$ . Since the barycenter is an affine function (Exercise 9.6.2), it follows that

$$\text{bar}(W) = \sum_{P \in \mathcal{P}_n} W(P) \text{bar}(W_P)$$

and, therefore,

$$H(\text{bar}(W)) = \sum_{P \in \mathcal{P}_n} W(P)H(\text{bar}(W_P)).$$

Define  $H_n(\eta) = H(\text{bar}(W_{\mathcal{P}_n(\eta)}))$ , for each  $\eta \in \mathcal{M}$ . Then the last identity above may be rewritten as follows:

$$H(\text{bar}(W)) = \int H_n(\eta) dW(\eta) \quad \text{for every } n. \tag{9.6.10}$$

It follows directly from the definition of  $H_n$  that  $0 \leq H_n(\eta) \leq \sup H$  for every  $n$  and every  $\eta$ . Recall that  $\sup H < \infty$  (Corollary 9.6.9). We also claim that

$$\limsup_n H_n(\eta) \leq H(\eta) \quad \text{for every } \eta \in \mathcal{M}. \tag{9.6.11}$$

This may be seen as follows. Given any neighborhood  $V = V(\eta, \Phi, \varepsilon)$  of  $\eta$ , we have that  $\mathcal{P}_n(\eta) \subset V$  for every large  $n$ , because the diameter of  $\mathcal{P}_n(\eta)$  converges to zero. Then, assuming always that  $W(\mathcal{P}_n(\eta))$  is positive,

$$W_{\mathcal{P}_n(\eta)}(V) \geq W_{\mathcal{P}_n(\eta)}(\mathcal{P}_n(\eta)) = 1.$$

By Lemma 9.6.7, it follows that  $\text{bar}(W_{\mathcal{P}_n(\eta)}) \in V$  for every large  $n$ . Now (9.6.11) is a direct consequence of the hypothesis that  $H$  is upper semi-continuous.

Applying the lemma of Fatou to the sequence  $-H_n + \sup H$ , we deduce that

$$\limsup_n \int H_n(\eta) dW(\eta) \leq \int \limsup_n H_n(\eta) dW(\eta) \leq \int H(\eta) dW(\eta). \tag{9.6.12}$$

Combining the relations (9.6.10) and (9.6.12), we get that

$$H(\text{bar}(W)) \leq \int H(\eta) dW(\eta),$$

as we wanted to prove.

Now the proof of Theorems 9.6.2 and 9.6.5 is complete.

### 9.6.3 Exercises

- 9.6.1. Check that, given any probability measure  $W$  on  $\mathcal{M}_1(M)$ , there exists a unique probability measure  $\text{bar}(W) \in \mathcal{M}_1(M)$  on  $M$  that satisfies (9.6.4).
- 9.6.2. Show that the barycenter function is strongly affine: if  $W_i, i \geq 1$  are probability measures on  $\mathcal{M}_1(M)$  and  $t_i, i \geq 1$  are non-negative numbers with  $\sum_{i=1}^\infty t_i = 1$ , then

$$\text{bar}\left(\sum_{i=1}^\infty t_i W_i\right) = \sum_{i=1}^\infty t_i \text{bar}(W_i).$$

- 9.6.3. Show that if  $\mathcal{M} \subset \mathcal{M}_1(M)$  is a closed convex set then  $\mathcal{M}$  is strongly convex. Moreover, in that case  $W(\mathcal{M}) = 1$  implies that  $\text{bar}(W) \in \mathcal{M}$ .
- 9.6.4. Check that the inequality  $\geq$  in Theorem 9.6.2 may also be obtained through the following, more direct, argument:
  - (1) Recalling that the function  $\phi(x) = -x \log x$  is concave, show that  $H_\mu(\mathcal{Q}) \geq \int H_{\mu_P}(\mathcal{Q}) d\hat{\mu}(P)$  for every finite partition  $\mathcal{Q}$ .
  - (2) Deduce that  $h_\mu(f, \mathcal{Q}) \geq \int h_{\mu_P}(f, \mathcal{Q}) d\hat{\mu}(P)$  for every finite partition  $\mathcal{Q}$ .
  - (3) Conclude that  $h_\mu(f) \geq \int h_{\mu_P}(f) d\hat{\mu}(P)$ .

- 9.6.5. The inequality  $\leq$  in Theorem 9.6.2 is based on the fact that  $h_\mu(f, \mathcal{Q}) \leq \int h_{\mu_P}(f, \mathcal{Q}) d\hat{\mu}(P)$  for every finite partition  $\mathcal{Q}$ , which is part of Lemma 9.6.6. Find what is wrong with the following ‘‘alternative proof’’:

Let  $\mathcal{Q}$  be a finite partition. The theorem of Shannon–McMillan–Breiman ensures that  $h_\mu(f, \mathcal{Q}) = \int h_\mu(f, \mathcal{Q}, x) d\mu(x)$ , where

$$h_\mu(f, \mathcal{Q}, x) = \lim_n -\frac{1}{n} \log \mu(\mathcal{Q}^n(x)) = \lim_n -\frac{1}{n} \log \int \mu_P(\mathcal{Q}^n(x)) d\hat{\mu}(P).$$

By the Jensen inequality applied to the convex function  $\psi(x) = -\log x$ ,

$$\lim_n -\frac{1}{n} \log \int \mu_P(\mathcal{Q}^n(x)) d\hat{\mu}(P) \leq \lim_n \int -\frac{1}{n} \log \mu_P(\mathcal{Q}^n(x)) d\hat{\mu}(P).$$

Using the fact that  $h_{\mu_P}(f, \mathcal{Q}) = h_{\mu_P}(f, \mathcal{Q}, x)$  at almost every point (because the measure  $\mu_P$  is ergodic),

$$\begin{aligned} \lim_n \int -\frac{1}{n} \log \mu_P(\mathcal{Q}^n(x)) d\hat{\mu}(P) &= \int \lim_n -\frac{1}{n} \log \mu_P(\mathcal{Q}^n(x)) d\hat{\mu}(P) \\ &= \int h_{\mu_P}(f, \mathcal{Q}) d\hat{\mu}(P). \end{aligned}$$

This shows that  $h_{\mu}(f, \mathcal{Q}, x) \leq \int h_{\mu_P}(f, \mathcal{Q}) d\hat{\mu}(P)$  for every finite partition  $\mathcal{Q}$  and  $\mu$ -almost every  $x$ . Consequently,  $h_{\mu}(f, \mathcal{Q}) \leq \int h_{\mu_P}(f, \mathcal{Q}) d\hat{\mu}(P)$  for every finite partition  $\mathcal{Q}$ .

### 9.7 Jacobians and the Rokhlin formula

Let  $U$  be an open subset and  $m$  be the Lebesgue measure of  $\mathbb{R}^d$ . Let  $f : U \rightarrow U$  be a local diffeomorphism. By the formula of change of variables,

$$m(f(A)) = \int_A |\det Df(x)| dx \tag{9.7.1}$$

for any measurable subset  $A$  of a small ball restricted to which  $f$  is injective. The notion of a Jacobian that we present in this section extends this kind of relation to much more general transformations and measures. Besides introducing this concept, we are going to show that Jacobians do exist under quite general hypotheses. Most important, it is possible to express the system's entropy explicitly in terms of the Jacobian. Actually, we already encountered an interesting manifestation of this fact in Proposition 9.4.2.

Let  $f : M \rightarrow M$  be a measurable transformation. We say that  $f$  is *locally invertible* if there exists some countable cover  $\{U_k : U_k \geq 1\}$  of  $M$  by measurable sets such that  $f(U_k)$  is a measurable set and the restriction  $f|_{U_k} : U_k \rightarrow f(U_k)$  is a bijection with measurable inverse, for every  $k \geq 1$ . Every measurable subset of some  $U_k$  is called an *invertibility domain*. Note that the image  $f(A)$  of any invertibility domain  $A$  is a measurable set. It is also clear that if  $f$  is locally invertible then the pre-image  $f^{-1}(y)$  of any  $y \in M$  is countable: it contains at most one point in each  $U_k$ .

Let  $\eta$  be a probability measure on  $M$ , not necessarily invariant under  $f$ . A measurable function  $\zeta : M \rightarrow [0, \infty)$  is a *Jacobian* of  $f$  with respect to  $\eta$  if the restriction of  $\zeta$  to any invertibility domain  $A$  is integrable with respect to  $\eta$  and satisfies

$$\eta(f(A)) = \int_A \zeta d\eta. \tag{9.7.2}$$

It is important to note (see Exercise 9.7.1) that the definition does not depend on the choice of the family  $\{U_k : k \geq 1\}$ .

**Example 9.7.1.** Let  $\sigma : \Sigma \rightarrow \Sigma$  be the shift map in  $\Sigma = \{1, 2, \dots, d\}^{\mathbb{N}}$  and  $\mu$  be the Bernoulli measure associated with a probability vector  $p = (p_1, \dots, p_d)$ . The restriction of  $\sigma$  to each cylinder  $[0; a]$  is an invertible map. Moreover, given any cylinder  $[0; a, a_1, \dots, a_n] \subset [0; a]$ ,

$$\mu(\sigma([0; a, a_1, \dots, a_n])) = p_{a_1} \cdots p_{a_n} = \frac{1}{p_a} \mu([0; a, a_1, \dots, a_n]).$$

We invite the reader to deduce that  $\mu(\sigma(A)) = (1/p_a)\mu(A)$  for every measurable set  $A \subset [0; a]$ . Therefore, the function  $\zeta((x_n)_n) = 1/p_{x_0}$  is a Jacobian of  $\sigma$  with respect to  $\mu$ .

We say that a measure  $\eta$  is *non-singular* with respect to the transformation  $f : M \rightarrow M$  if the image of any invertibility domain with measure zero also has measure zero: if  $\eta(A) = 0$  then  $\eta(f(A)) = 0$ . For example, if  $f : U \rightarrow U$  is a local diffeomorphism of an open subset of  $\mathbb{R}^d$  and  $\eta$  is the Lebesgue measure, then  $\eta$  is non-singular. For any locally invertible transformation, every invariant probability measure is non-singular restricted to some full measure invariant set (Exercise 9.7.8).

It follows immediately from the definition (9.7.2) that if  $f$  admits a Jacobian with respect to a measure  $\eta$  then this measure is non-singular. The converse is also true:

**Proposition 9.7.2.** *Let  $f : M \rightarrow M$  be a locally invertible transformation and  $\eta$  be a measure on  $M$ , non-singular with respect to  $f$ . Then there exists some Jacobian of  $f$  with respect to  $\eta$  and it is essentially unique: any two Jacobians coincide at  $\eta$ -almost every point.*

*Proof.* We start by proving existence. Given any countable cover  $\{U_k : k \geq 1\}$  of  $M$  by invertibility domains of  $f$ , define  $P_1 = U_1$  and  $P_k = U_k \setminus (U_1 \cup \dots \cup U_{k-1})$  for each  $k > 1$ . Then  $\mathcal{P} = \{P_k : k \geq 1\}$  is a partition of  $M$  formed by invertibility domains. For each  $P_k \in \mathcal{P}$ , denote by  $\eta_k$  the measure defined on  $P_k$  by  $\eta_k(A) = \eta(f(A))$ . Equivalently,  $\eta_k$  is the image under  $(f|P_k)^{-1}$  of the measure  $\eta$  restricted to  $f(P_k)$ . The hypothesis that  $\eta$  is non-singular implies that every  $\eta_k$  is absolutely continuous with respect to  $\eta$  restricted to  $P_k$ :

$$\eta(A) = 0 \quad \Rightarrow \quad \eta_k(A) = \eta(f(A)) = 0$$

for every measurable set  $A \subset P_k$ . Let  $\zeta_k = d\eta_k/d(\eta|P_k)$  be the Radon–Nikodym derivative (Theorem A.2.18). Then,  $\zeta_k$  is a function defined on  $P_k$ , integrable with respect to  $\eta$  and satisfying

$$\eta(f(A)) = \eta_k(A) = \int_A \zeta_k d\eta \tag{9.7.3}$$

for every measurable set  $A \subset P_k$ . Consider the function  $\zeta : M \rightarrow [0, \infty)$  whose restriction to each  $P_k \in \mathcal{P}$  is given by  $\zeta_k$ . Every subset of  $U_k$  may be written

as a (disjoint) union of subsets of  $P_1, \dots, P_k$ . Applying (9.7.3) to each one of these subsets and summing the corresponding equalities, we get that

$$\eta(f(A)) = \int_A \xi \, d\eta \quad \text{for every measurable set } A \subset U_k \text{ and } k \geq 1.$$

This proves that  $\xi$  is a Jacobian of  $f$  with respect to  $\eta$ .

Now suppose that  $\xi$  and  $\zeta$  are Jacobians of  $f$  with respect to  $\eta$  and there exists  $B \subset M$  with  $\eta(B) > 0$  such that  $\xi(x) \neq \zeta(x)$  for every  $x \in B$ . Up to replacing  $B$  by a suitable subset, and exchanging the roles of  $\xi$  and  $\zeta$  if necessary, we may suppose that  $\xi(x) < \zeta(x)$  for every  $x \in B$ . Similarly, we may suppose that  $B$  is contained in some  $U_k$ . Then,

$$\eta(f(B)) = \int_B \xi \, d\eta < \int_B \zeta \, d\eta = \eta(f(B)).$$

This contradiction proves that the Jacobian is essentially unique.

From now on, we denote by  $J_\eta f$  the (essentially unique) Jacobian of a locally invertible transformation  $f : M \rightarrow M$  with respect to a measure  $\eta$ , when it exists.

By definition,  $J_\eta f$  is integrable on each invertibility domain. If  $f$  is such that the number of pre-images of any  $y \in M$  is bounded then the Jacobian is (globally) integrable: if  $\ell \geq 1$  is the maximum number of pre-images then

$$\int J_\eta f \, d\eta = \sum_k \int_{P_k} J_\eta f \, d\eta = \sum_k \eta(f(P_k)) \leq \ell,$$

because every point  $y \in M$  is in no more than  $\ell$  images  $f(P_k)$ .

The following observation will be useful in the sequel. Let  $Z \subset M$  be the set of points where the Jacobian  $J_\eta f$  vanishes. Covering  $Z$  with a countable family of invertibility domains and using (9.7.2), we see that  $f(Z)$  is a measurable set and  $\eta(f(Z)) = 0$ . In other words, the set of points  $y \in M$  such that  $J_\mu f(x) > 0$  for every  $x \in f^{-1}(y)$  has total measure for  $\eta$ . When the probability measure  $\eta$  is invariant under  $f$ , it also follows that  $\eta(f^{-1}(f(Z))) = \eta(f(Z)) = 0$  and so  $\eta(Z) = 0$ .

The main result in this section is the following formula for the entropy of an invariant measure:

**Theorem 9.7.3** (Rokhlin formula). *Let  $f : M \rightarrow M$  be a locally invertible transformation and  $\mu$  be a probability measure invariant under  $f$ . Assume that there is some partition  $\mathcal{P}$  with finite entropy such that  $\bigcup_n \mathcal{P}^n$  generates the  $\sigma$ -algebra of  $M$ , up to measure zero, and every  $P \in \mathcal{P}$  is an invertibility domain of  $f$ . Then  $h_\mu(f) = \int \log J_\mu f \, d\mu$ .*

*Proof.* Let us consider the sequence of partitions  $\mathcal{Q}_n = \bigvee_{j=1}^n f^{-j}(\mathcal{P})$ . By Corollary 9.2.5 and Lemma 9.1.12,

$$h_\mu(f) = h_\mu(f, \mathcal{P}) = \lim_n H_\mu(\mathcal{P}/\mathcal{Q}_n). \tag{9.7.4}$$

By definition (as before,  $\phi(x) = -x \log x$ ),

$$\begin{aligned} H_\mu(\mathcal{P}/\mathcal{Q}_n) &= \sum_{P \in \mathcal{P}} \sum_{Q_n \in \mathcal{Q}_n} -\mu(P \cap Q_n) \log \frac{\mu(P \cap Q_n)}{\mu(Q_n)} \\ &= \sum_{P \in \mathcal{P}} \sum_{Q_n \in \mathcal{Q}_n} \mu(Q_n) \phi\left(\frac{\mu(P \cap Q_n)}{\mu(Q_n)}\right). \end{aligned} \tag{9.7.5}$$

Let  $e_n(\psi, x)$  be the conditional expectation of a function  $\psi$  with respect to the partition  $\mathcal{Q}_n$  and  $e(\psi, x)$  be its limit when  $n$  goes to infinity (these notions were introduced in Section 5.2.1: see (5.2.1) and Lemma 5.2.1). It is clear from the definition that

$$\frac{\mu(P \cap Q_n)}{\mu(Q_n)} = e_n(\mathcal{X}_P, x) \quad \text{for every } x \in Q_n \text{ and every } Q_n \in \mathcal{Q}_n.$$

Therefore,

$$\sum_{P \in \mathcal{P}} \sum_{Q_n \in \mathcal{Q}_n} \mu(Q_n) \phi\left(\frac{\mu(P \cap Q_n)}{\mu(Q_n)}\right) = \sum_{P \in \mathcal{P}} \int \phi(e_n(\mathcal{X}_P, x)) d\mu(x). \tag{9.7.6}$$

By Lemma 5.2.1, the limit  $e(\mathcal{X}_P, x) = \lim_n e_n(\mathcal{X}_P, x)$  exists at  $\mu$ -almost every  $x$ . So, observing that the function  $\phi$  is bounded, we may use the dominated convergence theorem to deduce from (9.7.4)–(9.7.6) that

$$h_\mu(f) = \sum_{P \in \mathcal{P}} \int \phi(e(\mathcal{X}_P, x)) d\mu(x). \tag{9.7.7}$$

Now we need to relate the expression inside the integral to the Jacobian. This we do by means of Lemma 9.7.5 below. Beforehand, let us prove the following *change of variables formulas*:

**Lemma 9.7.4.** *For any probability measure  $\eta$  non-singular with respect to  $f$ , and any invertibility domain  $A \subset M$  of  $f$ :*

- (i)  $\int_{f(A)} \varphi d\eta = \int_A (\varphi \circ f) J_\eta f d\eta$  for any measurable function  $\varphi : f(A) \rightarrow \mathbb{R}$  such that the integrals are defined (possibly  $\pm\infty$ ).
- (ii)  $\int_A \psi d\eta = \int_{f(A)} (\psi / J_\eta f) \circ (f|_A)^{-1} d\eta$  for any measurable function  $\psi : A \rightarrow \mathbb{R}$  such that the integrals are defined (possibly  $\pm\infty$ ).

*Proof.* The definition (9.7.2) means that the formula in part (i) holds for the characteristic function  $\varphi = \mathcal{X}_{f(A)}$  for any invertibility domain  $A$ . Thus, it holds for the characteristic function of any measurable subset of  $f(A)$ , since such a subset may be written as  $f(B)$  for some invertibility domain  $B \subset A$ . Hence, by linearity, the identity extends to every simple function defined on  $f(A)$ . Using the monotone convergence theorem, we conclude that the identity holds for every non-negative measurable function. Using linearity once more, we get the general statement of part (i).



To deduce the claim in (ii), apply (i) to the function  $\varphi = (\psi/J_\eta f) \circ (f|A)^{-1}$ . Note that this function is well defined at  $\eta$ -almost every point for, as observed before,  $J_\eta f(x) > 0$  for every  $x$  in the pre-image of  $\eta$ -almost every  $y \in M$ .

**Lemma 9.7.5.** *For every bounded measurable function  $\psi : M \rightarrow \mathbb{R}$  and every probability measure  $\eta$  invariant under  $f$ ,*

$$e(\psi, x) = \hat{\psi}(f(x)) \text{ for } \eta\text{-almost every } x, \text{ where } \hat{\psi}(y) = \sum_{z \in f^{-1}(y)} \frac{\psi}{J_\eta f}(z).$$

*Proof.* Recall that  $\mathcal{Q}_n = \bigvee_{j=1}^n f^{-j}(\mathcal{P})$ , that is,  $\mathcal{Q}_n(x) = \bigcap_{j=1}^n f^{-1}(\mathcal{P}(f^j(x)))$  for each  $x$ . We also use the sequence of partitions  $\mathcal{P}^n = \bigvee_{j=0}^{n-1} f^{-j}(\mathcal{P})$ . Observe that  $\mathcal{Q}_n(x) = f^{-1}(\mathcal{P}^{n-1}(f(x)))$  and  $\mathcal{P}^n(x) = \mathcal{P}(x) \cap \mathcal{Q}_n(x)$  for every  $n$  and every  $x$ . Then,

$$\int_{\mathcal{P}^{n-1}(f(x))} \hat{\psi} \, d\eta = \sum_{P \in \mathcal{P}} \int_{f(P) \cap \mathcal{P}^{n-1}(f(x))} \frac{\psi}{J_\eta f} \circ (f|P)^{-1} \, d\eta.$$

Using the formula of change of variables in Lemma 9.7.4(ii), the expression on the right-hand side may be rewritten as

$$\sum_{P \in \mathcal{P}} \int_{P \cap \mathcal{Q}_n(x)} \psi(z) \, d\eta(z) = \int_{\mathcal{Q}_n(x)} \psi \, d\eta.$$

Therefore,

$$\int_{\mathcal{P}^{n-1}(f(x))} \hat{\psi} \, d\eta = \int_{\mathcal{Q}_n(x)} \psi \, d\eta. \tag{9.7.8}$$

Let  $e'_{n-1}(\hat{\psi}, x)$  be the conditional expectation of  $\hat{\psi}$  with respect to the partition  $\mathcal{P}^{n-1}$ , as defined in Section 5.2.1, and let  $e'(\hat{\psi}, x)$  be its limit when  $n$  goes to infinity, given by Lemma 5.2.1. The hypothesis that  $\eta$  is invariant gives that  $\eta(\mathcal{P}^{n-1}(f(x))) = \eta(\mathcal{Q}_n(x))$ . Dividing both sides of (9.7.8) by this number, we get that

$$e'_{n-1}(\hat{\psi}, f(x)) = e_n(\psi, x) \text{ for every } x \text{ and every } n > 1. \tag{9.7.9}$$

Then, taking the limit,  $e'(\hat{\psi}, f(x)) = e(\psi, x)$  for  $\eta$ -almost every  $x$ . On the other hand, according to Exercise 5.2.3, the hypothesis implies that  $e'(\hat{\psi}, y) = \hat{\psi}(y)$  for  $\eta$ -almost every  $y \in M$ .

Let us apply this lemma to  $\psi = \mathcal{X}_P$  and  $\eta = \mu$ . Since  $f$  is injective on every element of  $\mathcal{P}$ , each intersection  $P \cap f^{-1}(y)$  either is empty or contains exactly one point. Therefore, it follows from Lemma 9.7.5 that  $e(\mathcal{X}_P, x) = \hat{\mathcal{X}}_P(f(x))$ , with

$$\hat{\mathcal{X}}_P(y) = \begin{cases} 1/J_{\mu} f((f|P)^{-1}(y)) & \text{if } y \in f(P) \\ 0 & \text{if } y \notin f(P). \end{cases}$$

Then, recalling that the measure  $\mu$  is assumed to be invariant,

$$\begin{aligned} \int \phi(e(\mathcal{X}_P, x)) d\mu(x) &= \int \phi(\hat{\mathcal{X}}_P(y)) d\mu(y) \\ &= \int_{f(P)} \left( \frac{1}{J_{\mu}f} \log J_{\mu}f \right) \circ (f|_P)^{-1} d\mu = \int_P \log J_{\mu}f d\mu \end{aligned}$$

(the last step uses the identity in part (ii) of Lemma 9.7.4). Replacing this expression in (9.7.7), we get that

$$h_{\mu}(f) = \sum_{P \in \mathcal{P}} \int_P \log J_{\mu}f d\mu = \int \log J_{\mu}f d\mu,$$

as stated in the theorem.

### 9.7.1 Exercises

- 9.7.1. Check that the definition of a Jacobian does not depend on the choice of the cover  $\{U_k : k \geq 1\}$  by invertibility domains.
- 9.7.2. Let  $\sigma : \Sigma \rightarrow \Sigma$  be the shift map in  $\Sigma = \{1, 2, \dots, d\}^{\mathbb{N}}$  and  $\mu$  be the Markov measure associated with an aperiodic matrix  $P$ . Find the Jacobian of  $f$  with respect to  $\mu$ .
- 9.7.3. Let  $f : M \rightarrow M$  be a locally invertible transformation and  $\eta$  be a probability measure on  $M$ , non-singular with respect to  $f$ . Show that for every bounded measurable function  $\psi : M \rightarrow \mathbb{R}$ ,

$$\int \psi d\eta = \int \sum_{z \in f^{-1}(x)} \frac{\psi}{J_{\eta}f}(z) d\eta(x).$$

- 9.7.4. Let  $f : M \rightarrow M$  be a locally invertible transformation and  $\eta$  be a probability measure on  $M$ , non-singular with respect to  $f$ . Show that  $\eta$  is invariant under  $f$  if and only if

$$\sum_{z \in f^{-1}(x)} \frac{1}{J_{\eta}f}(z) = 1 \quad \text{for } \eta\text{-almost every } x \in M.$$

Moreover, if  $\eta$  is invariant under  $f$  then  $J_{\eta}f \geq 1$  at  $\mu$ -almost every point.

- 9.7.5. Let  $f : M \rightarrow M$  be a locally invertible transformation and  $\eta$  be a probability measure on  $M$ , non-singular with respect to  $f$ . Show that, for every  $k \geq 1$ , there exists a Jacobian of  $f^k$  with respect to  $\eta$  and it is given by

$$J_{\eta}f^k(x) = \prod_{j=0}^{k-1} J_{\eta}f(f^j(x)) \quad \text{for } \eta\text{-almost every } x.$$

Assuming that  $f$  is invertible, what can be said about the Jacobian of  $f^{-1}$  with respect to  $\eta$ ?

- 9.7.6. Let  $f : M \rightarrow M$  and  $g : N \rightarrow N$  be locally invertible transformations and let  $\mu$  and  $\nu$  be probability measures invariant under  $f$  and  $g$ , respectively. Assume that there exists an ergodic equivalence  $\phi : M \rightarrow N$  between the systems  $(f, \mu)$  and  $(g, \nu)$ . Show that  $J_{\mu}f = J_{\nu}g \circ \phi$  at  $\mu$ -almost every point.

- 9.7.7. Let  $\sigma_k : \Sigma_k \rightarrow \Sigma_k$  and  $\sigma_l : \Sigma_l \rightarrow \Sigma_l$  be the shift maps in  $\Sigma_k = \{1, \dots, k\}^{\mathbb{N}}$  and  $\Sigma_l = \{1, \dots, l\}^{\mathbb{N}}$ . Let  $\mu_k$  and  $\mu_l$  be the Bernoulli measures on  $\Sigma_k$  and  $\Sigma_l$ , respectively, associated with probability vectors  $p = (p_1, \dots, p_k)$  and  $q = (q_1, \dots, q_l)$ . Show that the systems  $(\sigma_k, \mu_k)$  and  $(\sigma_l, \mu_l)$  are ergodically equivalent if and only if  $k = l$  and the vectors  $p$  and  $q$  are obtained from one another by permutation of the components.
- 9.7.8. Let  $\mu$  be a probability measure invariant under a locally invertible transformation  $f : M \rightarrow M$ . Show that there exists a full measure set  $N \subset M$  such that  $N \subset f^{-1}(N)$  and  $\mu$  restricted to  $N$  is non-singular with respect to the restriction  $f : N \rightarrow N$ . Conclude that  $f$  admits a Jacobian with respect to  $\mu$ .