1

Measure-Theoretic Entropy

Measure-theoretic entropy is a numerical invariant associated to a measure-preserving system. The early part of the theory described here is due essentially to Kolmogorov, Sinaĭ and Rokhlin, and dates[(1)] from the late 1950s.

Entropy will be used in several ways, most particularly in order to distinguish Haar measures, or other homogeneous measures, from other invariant measures. One of the initial motivations for this theory was the following kind of question. The Bernoulli shift on 2 symbols

$$\sigma_{(2)} : \{0,1\}^{\mathbb{Z}} \to \{0,1\}^{\mathbb{Z}}$$

preserving the $(\frac{1}{2}, \frac{1}{2})$ Bernoulli measure μ_2, and the Bernoulli shift on 3 symbols

$$\sigma_{(3)} : \{0,1,2\}^{\mathbb{Z}} \to \{0,1,2\}^{\mathbb{Z}}$$

preserving the $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ Bernoulli measure μ_3, share many properties, and in particular are unitarily equivalent*. Are they isomorphic as measure-preserving transformations? To see that this is not out of the question, notice that Mešalkin [96] showed that the $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ Bernoulli shift is isomorphic to the one defined by the probability vector $(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$. A brief description of the isomorphism between these two maps is given in Section 1.6; see also the book of Cornfeld, Fomin and Sinaĭ [27, Sect. 8.1] and the survey article of Weiss [145, Sect. 5]. It turns out that entropy is preserved by measurable isomorphism, and the $(\frac{1}{2}, \frac{1}{2})$ Bernoulli shift and the $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ Bernoulli shift have different entropies and so they cannot be isomorphic.

The basic machinery of entropy theory will take some effort to develop, but once the foundations are in place only the main properties will be needed.

* That is, there is an invertible linear operator $W : L^2_{\mu_3} \to L^2_{\mu_2}$ with

$$\langle Wf, Wg \rangle_{\mu_2} = \langle f, g \rangle_{\mu_3}$$

and $U_{\sigma_2} = W U_{\sigma_3} W^{-1}$.

1.1 Entropy of a Partition

Recall that a *partition** of a probability space $(X, \mathscr{B}, \mu)$ is a finite or countably infinite collection of disjoint (and, by assumption, always) measurable subsets of X whose union is X,

$$\xi = \{A_1, \dots, A_k\} \text{ or } \xi = \{A_1, A_2, \dots\}.$$

For any partition ξ we define $\sigma(\xi)$ to be the smallest σ-algebra containing the elements of ξ. We will call the elements of ξ the *atoms* of the partition, and write $[x]_\xi$ for the atom of ξ containing x. If the partition ξ is finite, then the σ-algebra $\sigma(\xi)$ is also finite and comprises the unions of elements of ξ.

If ξ and η are partitions, then $\xi \leqslant \eta$ means that each atom of ξ is a union of atoms of η, or η is a *refinement* of ξ. The *common refinement* of

$$\xi = \{A_1, A_2, \dots\}$$

and

$$\eta = \{B_1, B_2, \dots\},$$

denoted $\xi \vee \eta$, is the partition into all sets of the form $A_i \cap B_j$.

Notice that $\sigma(\xi \vee \eta) = \sigma(\xi) \vee \sigma(\eta)$ where the right-hand side denotes the σ-algebra generated by $\sigma(\xi)$ and $\sigma(\eta)$, equivalently the intersection of all sub–σ–algebras of $\mathscr{B}$ containing both $\sigma(\xi)$ and $\sigma(\eta)$. This allows us to move from partitions to subalgebras with impunity. The notation $\bigvee_{n=0}^{\infty} \xi_n$ will always mean the smallest σ-algebra containing $\sigma(\xi_n)$ for all $n \geqslant 0$, and we will also write $\xi_n \nearrow \mathscr{B}$ as a shorthand for $\sigma(\xi_n) \nearrow \mathscr{B}$ for an increasing sequence of partitions that generate the σ-algebra $\mathscr{B}$ of X.

Now let $T : X \to X$ be a measurable map, and $\xi = \{A_1, A_2, \dots\}$ be a partition. Write $T^{-1}\xi$ for the partition $\{T^{-1}A_1, T^{-1}A_2, \dots\}$ obtained by taking pre-images.

1.1.1 Basic Definition

A partition $\xi = \{A_1, A_2, \dots\}$ may be thought of as giving the possible outcomes $1, 2, \dots$ of an experiment, with the probability of outcome i being $\mu(A_i)$. The first step is to associate a number $H(\xi)$ to ξ which describes the amount of uncertainty about the outcome of the experiment, or equivalently the amount of information gained by learning the outcome of the experiment. Two extremes are clear: if one of the sets A_i has $\mu(A_i) = 1$ then there is no uncertainty about the outcome, and no information to be gained by performing it,

* We will often think of a partition as being given with an explicit enumeration of its elements: that is, as a *list* of disjoint measurable sets that cover X. We will use the word 'partition' both for a collection of sets and for an enumerated list of sets. This is usually a matter of notational convenience, but in Sections 1.2, 1.4 and 1.6 it is essential that we work with an enumerated list.

so $H(\xi) = 0$. At the opposite extreme, if each atom A_i of a partition with k elements has $\mu(A_i) = \frac{1}{k}$, then we have maximal uncertainty about the outcome, and $H(\xi)$ should take on its maximum value (for given k) for such a partition.

Definition 1.1. *The* entropy *of a partition $\xi = \{A_1, A_2, \dots\}$ is*

$$H_\mu(\xi) = H(\mu(A_1), \dots) = -\sum_{i \geqslant 1} \mu(A_i) \log \mu(A_i) \in [0, \infty]$$

where $0 \log 0$ is defined to be 0. If $\xi = \{A_1, \dots\}$ and $\eta = \{B_1, \dots\}$ are partitions, then the conditional entropy *of the outcome of ξ once we have been told the outcome of η (briefly, the* conditional entropy of ξ given η*) is defined to be*

$$H_\mu(\xi|\eta) = \sum_{j=1}^{\infty} \mu(B_j) H\big(\frac{\mu(A_1 \cap B_j)}{\mu(B_j)}, \frac{\mu(A_2 \cap B_j)}{\mu(B_j)}, \dots\big). \tag{1.1}$$

The formula in equation (1.1) may be viewed as a weighted average of entropies of the partition ξ conditioned (that is, restricted to each atom and then normalized by the measure of that atom) on individual atoms $B_j \in \eta$.

Under the correspondence between partitions and σ-algebras, we may also view H_μ as being defined on any σ-algebra corresponding to a countably infinite or finite partition.

1.1.2 Essential Properties

Notice that the quantity $H_\mu(\xi)$ does not depend on the partition ξ, but only on the probability vector $(\mu(A_1), \mu(A_2), \dots)$. Restricting to finite probability vectors, H is defined on the space of finite-dimensional simplices

$$\Delta = \bigcup_k \Delta_k$$

where $\Delta_k = \{(p_1, \dots, p_k) \mid p_i \geqslant 0, \sum p_i = 1\}$, by

$$H(p_1, \dots, p_k) = -\sum_{i=1}^{k} p_i \log p_i.$$

Remarkably, the function in Definition 1.1 is essentially the only function obeying a natural set of properties reflecting the idea of quantifying the uncertainty about the outcome of an experiment. We now list some basic properties of $H_\mu(\cdot)$, $H(\cdot)$, and $H_\mu(\cdot|\cdot)$. Of these properties, (1) and (2) are immediate consequences of the definition, and (3) and (4) will be shown later.

(1) $H(p_1, \dots, p_k) \geqslant 0$, and $H(p_1, \dots, p_k) = 0$ if and only if some $p_i = 1$.
(2) $H(p_1, \dots, p_k, 0) = H(p_1, \dots, p_k)$.

(3) For each $k \geqslant 1$, H restricted to Δ_k is continuous, independent under permutation of the variables, and attains the maximum value $\log k$ at the point $(\frac{1}{k}, \ldots, \frac{1}{k})$.
(4) $H_\mu(\xi \vee \eta) = H_\mu(\eta) + H_\mu(\xi|\eta)$.

Khinchin [73, p. 9] showed that H_μ as defined in Definition 1.1 is the only function with these properties. In this chapter all these properties of the entropy function will be derived, but Khinchin's *characterization* of entropy in terms of the properties (1) to (4) above will not be used and will not be proved here.

1.1.3 Convexity

Many of the most fundamental properties of entropy are a consequence of convexity, and we now recall some elementary properties of convex functions.

Definition 1.2. *A function $\psi : (a, b) \to \mathbb{R}$ is* convex *if*

$$\psi\left(\sum_{i=1}^{n} t_i x_i\right) \leqslant \sum_{i=1}^{n} t_i \psi(x_i)$$

for all $x_i \in (a, b)$ and $t_i \in [0, 1]$ with $\sum_{i=1}^{n} t_i = 1$, and is strictly convex *if*

$$\psi\left(\sum_{i=1}^{n} t_i x_i\right) < \sum_{i=1}^{n} t_i \psi(x_i)$$

unless $x_i = x$ for some $x \in (a, b)$ and all i with $t_i > 0$.

Let us recall a simple consequence of this definition. Suppose that $a < s < t < u < b$. Then convexity of ψ implies that

$$\psi(t) = \psi\left(\frac{u-t}{u-s}s + \frac{t-s}{u-s}u\right) \leqslant \frac{u-t}{u-s}\psi(s) + \frac{t-s}{u-s}\psi(u),$$

which is equivalent to the inequality of slopes

$$\frac{\psi(t) - \psi(s)}{t-s} \leqslant \frac{\psi(u) - \psi(t)}{u-t}.$$

Lemma 1.3 (Jensen's inequality). *Let $\psi : (a, b) \to \mathbb{R}$ be a convex function and let $f : X \to (a, b)$ be a measurable function in L^1_μ on a probability space $(X, \mathscr{B}, \mu)$. Then*

$$\psi\left(\int f(x)\,\mathrm{d}\mu(x)\right) \leqslant \int \psi\left(f(x)\right)\,\mathrm{d}\mu(x). \tag{1.2}$$

If in addition ψ is strictly convex, then

$$\psi\left(\int f(x)\,\mathrm{d}\mu(x)\right) < \int \psi\left(f(x)\right)\,\mathrm{d}\mu(x) \tag{1.3}$$

unless $f(x) = t$ for μ-almost every $x \in X$ for some fixed $t \in (a, b)$.

In this lemma we permit $a = -\infty$ and $b = \infty$. Similar conclusions hold on half-open and closed intervals.

Proof of Lemma 1.3. Let $t = \int f \, \mathrm{d}\mu$, so that $t \in (a, b)$. Let

$$\beta = \sup_{a<s<t} \left\{ \frac{\psi(t) - \psi(s)}{t - s} \right\},$$

so that, by convexity,

$$\beta \leqslant \inf_{t<u<b} \left\{ \frac{\psi(u) - \psi(t)}{u - t} \right\}.$$

It follows that

$$\psi(s) \geqslant \psi(t) + \beta(s - t)$$

if $a < s < b$, so

$$\psi\left(f(x)\right) - \psi(t) - \beta\left(f(x) - t\right) \geqslant 0 \tag{1.4}$$

for every $x \in X$. Since ψ is continuous (a consequence of convexity), $x \mapsto \psi\left(f(x)\right)$ is measurable and we may integrate equation (1.4) to get

$$\int \psi \circ f \, \mathrm{d}\mu - \psi\left(\int f \, \mathrm{d}\mu\right) - \beta \int f \, \mathrm{d}\mu + \beta \int f \, \mathrm{d}\mu \geqslant 0,$$

showing equation (1.2).

If ψ is strictly convex, then $\psi(s) > \psi(t) + \beta(s - t)$ for all $s > t$ and for all $s < t$. If f is not equal almost everywhere to a constant, then $f(x) - t$ takes on both negative and positive values on sets of positive measure, proving equation (1.3). □

The shape of the graph of the function $x \mapsto x \log x$ determines further properties of the entropy function. Define a function $\phi : [0, \infty) \to \mathbb{R}$ by

$$\phi(x) = \begin{cases} 0 & \text{if } x = 0; \\ x \log x & \text{if } x > 0. \end{cases}$$

Clearly the choice of $\phi(0)$ means that ϕ is continuous at 0. The graph of ϕ is shown in Figure 1.1; the minimum value occurs at $x = 1/e$.

Since $\phi''(x) = \frac{1}{x} > 0$ and $(x \mapsto -\log x)'' = \frac{1}{x^2} > 0$ on $(0, 1]$, a simple calculation shows the following.

Lemma 1.4. *The function $x \mapsto \phi(x)$ is strictly convex on $[0, \infty)$ and the function $x \mapsto -\log x$ is strictly convex on $(0, \infty)$.*

A consequence of this is that the maximum amount of information in a partition arises when all the atoms of the partition have the same measure.

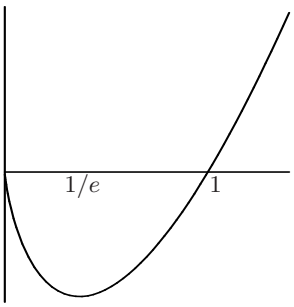

Fig. 1.1. The graph of $x \mapsto \phi(x)$.

Proposition 1.5. *If ξ is a partition with k atoms, then*

$$H_\mu(\xi) \leqslant \log k,$$

with equality if and only if $\mu(P) = \frac{1}{k}$ for each atom P of ξ.

This establishes property (3) of the function $H : \Delta \to [0, \infty)$ from p. 6. We also note that this proposition is a precursor of the kind of characterization of uniform (that is, Haar) measures as being those with maximal entropy (see Example 1.27 for the first instance of this phenomenon).

Proof of Proposition 1.5. By Lemma 1.4, if some atom P has

$$0 < \mu(P) \neq \tfrac{1}{k}$$

then

$$-\tfrac{1}{k} \log k = \phi\left(\tfrac{1}{k}\right) = \phi\Big(\sum_{P \in \xi} \tfrac{1}{k}\mu(P)\Big) < \sum_{P \in \xi} \tfrac{1}{k}\phi\left(\mu(P)\right),$$

so

$$-\sum_{P \in \xi} \mu(P) \log \mu(P) < \log k.$$

If $\mu(P) = \frac{1}{k}$ for all $P \in \xi$, then $H_\mu(\xi) = \log k$. □

1.1.4 Proof of Essential Properties

It will be useful to introduce a function associated to a partition ξ closely related to the entropy $H_\mu(\xi)$.

Definition 1.6. *The* information function *of a partition $\xi = \{A_1, A_1, \dots\}$ is defined by*

$$I_\mu(\xi)(x) = -\log \mu\left([x]_\xi\right),$$

where $[x]_\xi \in \xi$ is the partition element with $x \in [x]_\xi$. Moreover, if η is another partition, then the conditional information function *of ξ given η is defined by*

$$I_\mu\left(\xi\middle|\eta\right) = -\log \frac{\mu([x]_{\xi\vee\eta})}{\mu([x]_\eta)}.$$

A more sophisticated notion of entropy and of information function conditional on a given σ-algebra will be given in Definition 4.3. In the next proposition we give the remaining main properties of the entropy function, and in particular we prove property (4) from p. 6.

Proposition 1.7. *Let ξ and η be countable partitions of $(X, \mathscr{B}, \mu)$. Then*

(1) $H_\mu(\xi) = \int I_\mu(\xi)\,\mathrm{d}\mu$ *and* $H_\mu\left(\xi\middle|\eta\right) = \int I_\mu\left(\xi\middle|\eta\right)\,\mathrm{d}\mu$;

(2) $I_\mu(\xi \vee \eta) = I_\mu(\xi) + I_\mu(\eta\big|\xi)$, $H_\mu(\xi \vee \eta) = H_\mu(\xi) + H_\mu(\eta\big|\xi)$ *and so, if $H_\mu(\xi) < \infty$, then*

$$H_\mu(\eta\big|\,\xi) = H_\mu(\xi \vee \eta) - H_\mu(\xi);$$

(3) $H_\mu(\xi \vee \eta) \leqslant H_\mu(\xi) + H_\mu(\eta)$;

(4) *if η and ζ are partitions of finite entropy, then $H_\mu(\xi\big|\,\eta \vee \zeta) \leqslant H_\mu(\xi\big|\,\zeta)$.*

We note that all the properties in Proposition 1.7 fit very well with the interpretation of $I_\mu(\xi)(x)$ as the information gained about the point x by learning which atom of ξ contains x, and of $H_\mu(\xi)$ as the average information. Thus (2) says that the information gained by learning which element of the refinement $\xi\vee\eta$ contains x is equal to the information gained by learning which atom of ξ contains x added to the information gained by learning in addition which atom of η contains x *given* the earlier knowledge about which atom of ξ contains x. The reader may find it helpful to give similar interpretations of the various entropy and information identities and inequalities that come later.

Example 1.8. Notice that the relation $H_\mu(\xi_2\big|\xi_1) \leqslant H_\mu(\xi_2)$ for entropy (implied by properties (2) and (3)) does not hold for the information function $I_\mu\left(\cdot\middle|\cdot\right)$. For example, let ξ_1 and ξ_2 denote the partitions of $[0,1]^2$ shown in Figure 1.2, and let m denote the two-dimensional Lebesgue measure on $[0,1]^2$. Then $I_\mu(\xi_2) = \log 2$ while

$$I_\mu\left(\xi_2\middle|\xi_1\right) \text{ is } \begin{cases} > \log 2 \text{ in the shaded region;} \\ < \log 2 \text{ outside the shaded region.} \end{cases}$$

Proof of Proposition 1.7. Write $\xi = \{A_1, A_2, \dots\}$ and $\eta = \{B_1, B_2, \dots\}$; then

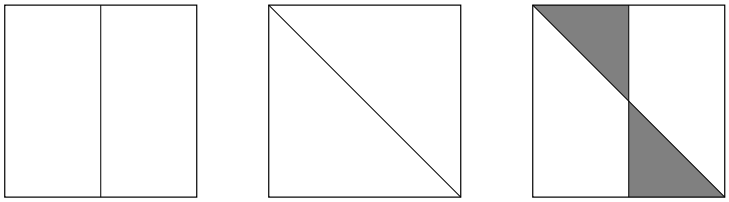

Fig. 1.2. Partitions ξ_1 and ξ_2 and their refinement.

$$\begin{aligned}\int I_\mu(\xi|\eta)\,\mathrm{d}\mu &= -\sum_{\substack{A_i\in\xi,\\ B_j\in\eta}} \left(\log\frac{\mu(A_i\cap B_j)}{\mu(B_j)}\right)\mu(A_i\cap B_j)\\ &= -\sum_{B_j\in\eta}\mu(B_j)\sum_{A_i\in\xi}\frac{\mu(A_i\cap B_j)}{\mu(B_j)}\log(A_i\cap B_j)\\ &= \sum_{B_j}\mu(B_j)H_\mu\left(\frac{\mu(A_1\cap B_j)}{\mu(B_j)},\dots\right)\\ &= H_\mu(\xi|\eta),\end{aligned}$$

showing the second formula in (1) and hence the first.

Notice that

$$\begin{aligned}I_\mu(\xi\vee\eta)(x) &= -\log\mu([x]_\xi\cap[x]_\eta)\\ &= -\log\mu([x]_\xi) - \log\frac{\mu([x]_\xi\cap[x]_\eta)}{\mu([x]_\xi)}\\ &= I_\mu(\xi)(x) + I_\mu(\eta|\xi)(x),\end{aligned}$$

which gives (2) by integration.

By convexity of ϕ,

$$\begin{aligned}H_\mu(\xi|\eta) &= -\sum_{A_i\in\xi}\sum_{B_j\in\eta}\phi\left(\frac{\mu(A_i\cap B_j)}{\mu(A_i)}\right)\\ &= -\sum_{B_j\in\eta}\sum_{A_i\in\xi}\mu(A_i)\phi\left(\frac{\mu(A_i\cap B_j)}{\mu(A_i)}\right)\\ &\leqslant \sum_{B_j\in\eta}\phi(B_j) = H_\mu(\xi),\end{aligned}$$

showing (3).

Finally,

$$\begin{aligned} H_\mu(\xi\big|\eta\vee\zeta) &= H_\mu(\xi\vee\eta\vee\zeta) - H_\mu(\eta\vee\zeta) \\ &= H_\mu(\xi) + H_\mu(\xi\vee\eta\big|\zeta) - H_\mu(\zeta) - H_\mu(\eta\big|\zeta) \\ &= \sum_{C\in\zeta}\mu(C)\left(H_{\mu(C)^{-1}\mu|_C}(\xi\vee\eta) - H_{\mu(C)^{-1}\mu|_C}(\eta)\right) \\ &= \sum_{C\in\zeta}\mu(C)H_{\mu(C)^{-1}\mu|_C}(\xi\big|\eta) \\ &\leqslant \sum_{C\in\zeta}\mu(C)H_{\mu(C)^{-1}\mu|_C}(\xi) = H_\mu(\xi\big|\eta). \end{aligned}$$

□

Exercises for Section 1.1

Exercise 1.1.1. Find countable partitions ξ, η of $[0,1]$ with $H_m(\xi) < \infty$ and $H_m(\eta) = \infty$, where m is Lebesgue measure.

Exercise 1.1.2. Show that the function

$$\mathsf{d}(\xi,\eta) = H_\mu(\xi\big|\eta) + H_\mu(\eta\big|\xi)$$

defines a metric on the space of all partitions of a probability space $(X, \mathscr{B}, \mu)$ with finite entropy.

Exercise 1.1.3. Two partitions ξ, η are independent, denoted $\xi \perp \eta$, if

$$\mu(A\cap B) = \mu(A)\mu(B)$$

for all $A \in \xi$ and $B \in \eta$. Prove that ξ and η with finite entropy are independent if and only if $H_\mu(\xi\vee\eta) = H_\mu(\xi) + H_\mu(\eta)$.

Exercise 1.1.4. For partitions ξ, η of fixed cardinality (and thought of as ordered lists), show that $H_\mu(\xi\big|\,\eta) = H_\mu(\xi\vee\eta) - H_\mu(\eta)$ is a continuous function of ξ and η with respect to the L^1_μ norm.

Exercise 1.1.5. Define sets

$$\Psi_k(X) = \{\text{partitions of } X \text{ with } k \text{ or fewer atoms}\},\ \ \Psi_{<\infty}(X) = \bigcup_{k\geqslant 1}\Psi_k$$

and

$$\Psi(X) = \{\text{partitions of } X \text{ with finite entropy}\}.$$

Prove that $\Psi_k(X)$ and $\Psi(X)$ are complete metric spaces under the partition metric from Exercise 1.1.2. Prove that $\Psi_{<\infty}(X)$ is dense in $\Psi(X)$.

1.2 Compression Algorithms

In this section we discuss a clearly related but slightly different point of view on the notions of information and entropy for finite or countable partitions[(2)]. It will be important here to think of a finite or countable partition $\xi = (A_1, A_2, \dots)$ as an ordered list rather than a set of subsets. We will refer to the indices $1, 2, \dots$ in the chosen enumeration of ξ as *symbols* or *source symbols* in the *alphabet*, which is a subset of $\mathbb{N}$.

We wish to encode each symbol by a finite binary sequence $d_1 \dots d_\ell$ of length $\ell \geqslant 1$ with $d_1, \dots, d_\ell \in \{0, 1\}$ with the following properties:

(1) every finite binary sequence is the code of at most one symbol

$$i \in \{1, 2, \dots\};$$

(2) if $d_1 \dots d_\ell$ is the code of some symbol then for every $k < \ell$ the binary sequence $d_1 \dots d_k$ is *not* the code of a symbol.

A *code* (because of the second condition, this is usually called a *prefix code*) is a map

$$s : \{1, 2, \dots\} \to \bigcup_{\ell \geqslant 1} \{0, 1\}^\ell$$

with these two properties.

These two properties allow the code to be decoded: given a code $d_1 \dots d_\ell$ the symbol encoded by the sequence can be deduced, and if the code is read from the beginning it is possible to work out when the whole sequence for that symbol has been read. Clearly the last requirement is essential if we want to successfully encode and decode not just a single symbol i but a list of symbols $w = i_0 i_1 \dots i_r$. We will call such a list of symbols a *name* in the alphabet $\{1, 2, \dots\}$. Because of the properties assumed for a code s we may extend the code from symbols to names by simply concatenating the codes of the symbols in the name to form one binary sequence $s(i_0)s(1_1)\dots s(i_r)$ without needing separators between the codes for individual symbols. The properties of the code mean that there is well-defined decoding map defined on the set of codes of names.

Example 1.9. (1) A simple example of a code defined on the alphabet $\{1, 2, 3\}$ is given by $s(1) = 0$, $s(2) = 10$, $s(3) = 11$. In this case the binary sequence 100011 is the code of the name 2113, because property (2) means that the sequence 100011 may be parsed into codes of symbols uniquely as $10\big|0\big|0\big|11 = s(2)s(1)s(1)s(3)$.
(2) Consider the set of all words appearing in a given dictionary. The goal of encoding names might be to find binary representations of sentences consisting of English words chosen from the dictionary appearing in this book.
(3) A possible code for the infinite alphabet $\{1, 2, 3, \cdots\}$ is given by

$$1 \longmapsto 10$$
$$2 \longmapsto 110$$
$$3 \longmapsto 1110$$

and so on. Clearly this also gives a code for any finite alphabet.

Given that there are many possible codes, a natural question is to ask for codes that are optimal with respect to some notion of weight or cost. To explore this we need additional structure, and in particular need to make assumptions about how frequently different symbols appear. Assume that every symbol has an assigned probability $v_i \in [0,1]$, so that $\sum_{i=1}^{\infty} v_i = 1$. In Example 1.9(2), we may think of v_i as the relative frequency of the English word represented by i in this book.

Let $|s(i)|$ denote the length of the codeword $s(i)$, and then the average length of the code is

$$L(s) = \sum_i v_i |s(i)|,$$

which may be finite or infinite depending on the code.

We wish to think of a code s as a compression algorithm, and in this viewpoint a code s is better (on average more efficient) than another code s' if the average length of the code s is smaller than the average length of the code s'. This allows us to give a new interpretation of the entropy of a partition in terms of the average length of an *optimal code* for a given distribution of relative frequencies.

Lemma 1.10. *For any code s the average length satisfies*

$$L(s) \log 2 \geqslant H(v_1, v_2, \dots) = -\sum_i v_i \log v_i.$$

In other words the entropy $H(v_1, v_2, \dots)$ of a probability vector $(v_1, v_2, \dots)$ gives a bound on the average effectiveness of any possible compression algorithm for the symbols $(1, 2, \dots)$ with relative frequencies $(v_1, v_2, \dots)$.

Proof of Lemma 1.10. We claim that the requirements on the code s imply Kraft's inequality[3]

$$\sum_i 2^{-|s(i)|} \leqslant 1. \tag{1.5}$$

To see this relation, interpret a binary sequence $d_1 \dots d_\ell$ as the address of the binary interval

$$I(d_1 \dots d_\ell) = \left(\tfrac{d_1}{2} + \tfrac{d_2}{2^2} + \dots + \tfrac{d_\ell}{2^\ell}, \tfrac{d_1}{2} + \tfrac{d_2}{2^2} + \dots + \tfrac{d_\ell + 1}{2^\ell}\right) \tag{1.6}$$

of length $\frac{1}{2^\ell}$. The requirements on the code mean precisely that all the intervals $I(s(i))$ for $i = 1, 2, \dots$ are disjoint, which proves equation (1.5). The lemma now follows by convexity of $x \mapsto -\log x$ (see Lemma 1.4):

$$\begin{aligned} L(s)\log 2 - H(v_1, v_2, \dots) &= \sum_i v_i |s(i)| \log 2 + \sum_i v_i \log v_i \\ &= -\sum_i v_i \log\left(\frac{2^{-|s(i)|}}{v_i}\right) \\ &\geqslant -\log \sum_i \frac{1}{2^{|s(i)|}} \geqslant 0. \end{aligned}$$

□

Lemma 1.10 and its proof suggest that there might always be a code that is as efficient as entropy considerations allow. As we show next, this is true if the algorithm is allowed a small amount of wastage.

Starting with the probability vector $(v_1, v_2, \dots)$ we may assume, by reordering if necessary, that $v_1 \geqslant v_2 \geqslant \cdots$. Define $\ell_i = \lceil -\log_2 v_1 \rceil$ (where $\lceil t \rceil$ denotes the smallest integer greater than or equal to t), so that ℓ_i is the smallest integer with $\frac{1}{2^{\ell_i}} \leqslant v_i$. Starting with the first digit, associate to $i = 1$ the interval $I_1 = \left(0, \frac{1}{2^{\ell_1}}\right)$, to $i = 2$ the interval $I_2 = \left(\frac{1}{2^{\ell_1}}, \frac{1}{2^{\ell_1}} + \frac{1}{2^{\ell_2}}\right)$, and in general associate to i the interval

$$I_i = \left(\sum_{j=1}^{i-1} \frac{1}{2^{\ell_j}}, \sum_{j=1}^{i} \frac{1}{2^{\ell_j}}\right)$$

of length $\frac{1}{2^{\ell_i}}$. We claim that every such interval is the interval $I(s(i))$ for a unique address $s(i) = d_1 \dots d_{|s(i)|}$ as in equation (1.6).

Lemma 1.11. *Every interval I_i as constructed above coincides with $I(s(i))$ for a binary sequence $s(i) = d_1 \cdots d_{|s(i)|}$ of length $|s(i)|$. This defines a code s with*

$$L(s)\log 2 \leqslant H(v_1, v_2, \dots) + \log 2.$$

That is, the entropy (divided by $\log 2$) is, to within one digit, the best possible average length of a code encoding the alphabet with the given probability vector describing its relative frequency distribution.

Proof of Lemma 1.11. The requirement that a binary interval $I = \left(\frac{a}{2^m}, \frac{b}{2^n}\right)$ with $a, b \in \mathbb{N}_0$ and $m, n \in \mathbb{N}$ has an address $d_1 \dots d_\ell$ in the sense of equation (1.6) is precisely the requirement that we can choose to represent the endpoints $\frac{a}{2^m}$ and $\frac{b}{2^n}$ in such a way that $\frac{a}{2^m} = \frac{a'}{2^\ell}$, $\frac{b}{2^n} = \frac{b'}{2^\ell}$ and $b' - a' = 1$. In other words, the length of the interval must be a power $\frac{1}{2^\ell}$ of 2 with $\ell \geqslant m, n$. The ordering of ξ chosen and the construction of the intervals ensures this property. It follows that every interval I_i constructed coincides with $I(s(i))$ for some binary sequence $s(i)$ of length $|s(i)|$. The disjointness of the intervals ensures that s is a code.

The average length of the code s is, by definition,

$$L(s) = \sum_i v_i |s(i)| = -\frac{1}{\log 2} \sum_i v_i \log \frac{1}{2^{|s(i)|}}$$
$$\leqslant -\frac{1}{\log 2} \sum_i v_i \log \left(\frac{v_i}{2}\right) = \frac{1}{\log 2} H(v_1, v_2, \dots) + 1.$$

□

Lemmas 1.10 and 1.11 together comprise the *source coding theorem* of information theory.

We may now also give an interpretation of the information function. Given the probability vector $(v_1, v_2, \dots)$ we see that $-\log_2 v_i$ (up to one digit) is the number of digits used to encode the symbol i in the nearly optimal code discussed above. Hence, we might say that the information function $\frac{1}{\log 2} I_\mu(\xi)$ gives the number of digits of the outcome of the experiment represented by ξ in the nearly optimal coding of all the possible outcomes of the experiment, where optimality is understood with respect to the probabilities $\mu(A_i)$ of the outcomes A_i of the experiment represented by $\xi = (A_1, A_2, \dots)$.

1.3 Entropy of a Measure-Preserving Transformation

In the last two sections we introduced and studied in some detail the notions of entropy and conditional entropy for partitions. In this section we start to apply this theory to the study of measure-preserving transformations, starting with the simple observation that such a transformation preserves conditional entropy in the following sense.

Lemma 1.12. *Let $(X, \mathscr{B}, \mu, T)$ be a measure-preserving system and let ξ, η be partitions. Then*

$$H_\mu\left(\xi \middle| \eta\right) = H_\mu\left(T^{-1}\xi \middle| T^{-1}\eta\right)$$

and

$$I_\mu\left(\xi \middle| \eta\right) \circ T = I_\mu\left(T^{-1}\xi \middle| T^{-1}\eta\right). \tag{1.7}$$

Proof. It is enough to show equation (1.7). Notice that $T^{-1}[Tx]_\eta = [x]_{T^{-1}\eta}$ for all x, so

$$\begin{aligned} I_\mu\left(\xi \middle| \eta\right)(Tx) &= -\log \frac{\mu([Tx]_\xi \cap [Tx]_\eta)}{\mu([Tx]_\eta} \\ &= -\log \frac{\mu([Tx]_\xi \cap [Tx]_\eta)}{\mu(T^{-1}[Tx]_\eta} \\ &= -\log \frac{\mu([Tx]_\xi \cap [Tx]_\eta)}{\mu([x]_{T^{-1}\eta}} \\ &= I_\mu\left(T^{-1}\xi \middle| T^{-1}\eta\right)(x) \end{aligned}$$

□

We are going to define the notion of entropy of a measure-preserving transformation; in order to do this a standard[(4)] result about convergence of subadditive sequences is needed.

Lemma 1.13 (Fekete). *Let (a_n) be a sequence of real numbers with the subadditive property*

$$a_{m+n} \leqslant a_m + a_n \text{ for all } m, n \geqslant 1.$$

Then $(\frac{1}{n}a_n)$ converges (possibly to $-\infty$), and

$$\lim_{n\to\infty} \frac{1}{n}a_n = \inf_{n\geqslant 1} \frac{1}{n}a_n.$$

Proof. Let $a = \inf_{n\in\mathbb{N}}\{\frac{a_n}{n}\}$, so $\frac{a_n}{n} \geqslant a$ for all $n \geqslant 1$. We assume here that $a > -\infty$ and leave the case $a = -\infty$ as an exercise. In our applications, we will always have $a_n \geqslant 0$ for all $n \geqslant 1$ and hence $a \geqslant 0$. Given $\varepsilon > 0$, pick $k \geqslant 1$ such that $\frac{a_k}{k} < a + \frac{1}{2}\varepsilon$. Now by the subadditive property, for any $m \geqslant 1$ and j, $0 \leqslant j < k$,

$$\begin{aligned}\frac{a_{mk+j}}{mk+j} &\leqslant \frac{a_{mk}}{mk+j} + \frac{a_j}{mk+j}\\ &\leqslant \frac{a_{mk}}{mk} + \frac{a_j}{mk}\\ &\leqslant \frac{ma_k}{mk} + \frac{ja_1}{mk}\\ &\leqslant \frac{a_k}{k} + \frac{a_1}{m} < a + \frac{1}{2}\varepsilon + \frac{a_1}{m}.\end{aligned}$$

So if $n = mk + j$ is large enough to ensure that $\frac{a_1}{m} < \frac{1}{2}\varepsilon$, then $\frac{a_n}{n} < a + \varepsilon$ as required. □

This simple lemma will be applied in the following way. Let T be a measure-preserving transformation of $(X, \mathscr{B}, \mu)$, and let ξ be a partition of X with finite entropy. Recall that we can think of ξ as an experiment with at most countably many possible outcomes, represented by the atoms of ξ. The entropy $H_\mu(\xi)$ measures the average amount of information conveyed about the points of the space by learning the outcome of this experiment. This quantity could be any non-negative number (or infinity) and of course has nothing to do with the transformation T.

If we think of $T : X \to X$ as representing evolution in time, then the partition $T^{-1}\xi$ corresponds to the same experiment one time unit later. In this sense the partition $\xi \vee T^{-1}\xi$ represents the joint outcome of the experiment ξ carried out now and in one unit of time, so $H_\mu(\xi \vee T^{-1}\xi)$ measures the average amount of information obtained by learning the outcome of the experiment applied twice in a row. If the partition $T^{-k}\xi$ is independent (see the definition in Exercise 1.1.3) of

$$\xi \vee T^{-1}\xi \vee \cdots \vee T^{-(k-1)}\xi$$

for all $k \geqslant 1$, then

$$H_\mu(\xi \vee T^{-1}\xi \vee \cdots \vee T^{-(n-1)}\xi) = H_\mu(\xi) + \cdots + H_\mu(T^{-(n-1)}\xi) = nH_\mu(\xi)$$

for all $n \geqslant 1$ by an induction using Exercise 1.1.3 and Lemma 1.12. In general, Proposition 1.7(3) and Lemma 1.12 show that

$$H_\mu(\xi \vee T^{-1}\xi \vee \cdots \vee T^{-(n-1)}\xi) \leqslant H_\mu(\xi) + \cdots + H_\mu(T^{-(n-1)}\xi) = nH_\mu(\xi),$$

so the quantity $H_\mu(\xi \vee T^{-1}\xi \vee \cdots \vee T^{-(n-1)}\xi)$ grows at most linearly in n. We shall see later that the asymptotic rate of this linear growth exists and is a significant invariant associated to T and ξ. This asymptotic rate will in general depend on the partition ξ, but once this dependence is eliminated the resulting rate is an invariant associated to T, the *(dynamical) entropy of T with respect to μ*. It is clear that the sequence (a_n) defined by

$$a_n = H_\mu(\xi \vee T^{-1}\xi \vee \cdots \vee T^{-(n-1)}\xi)$$

is subadditive in the sense of Lemma 1.13, which shows the claimed convergence in the next definition.

Definition 1.14. *Let $(X, \mathscr{B}, \mu, T)$ be a measure-preserving system and let ξ be a partition of X with finite entropy. Then the entropy of T with respect to ξ is*

$$h_\mu(T, \xi) = \lim_{n\to\infty} \frac{1}{n} H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right) = \inf_{n\geqslant 1} \frac{1}{n} H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right).$$

The entropy of T is

$$h_\mu(T) = \sup_{\xi : H_\mu(\xi) < \infty} h_\mu(T, \xi).$$

If $\mathscr{A}$ is a sub σ-algebra of $\mathscr{B}$ with $T^{-1}(\mathscr{A}) \subseteq \mathscr{A}$ then the conditional entropy of T given $\mathscr{A}$ is

$$h_\mu(T\big|\mathscr{A}) = \sup_{\xi : H_\mu(\xi) < \infty} h_\mu(T, \xi\big|\mathscr{A})$$

where

$$h_\mu(T, \xi\big|\mathscr{A}) = \lim_{n\to\infty} \frac{1}{n} H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}(\xi)\big|\mathscr{A}\right) = \inf_{n\geqslant 1} \frac{1}{n} H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}(\xi)\big|\mathscr{A}\right).$$

Example 1.15. Let $X_{(2)} = \{0,1\}^{\mathbb{Z}}$ with the Bernoulli $(\frac{1}{2}, \frac{1}{2})$ measure $\mu_{(2)}$, preserved by the shift σ_2. Consider the *state partition*

$$\xi = \{[0]_0, [1]_0\}$$

where $[0]_0 = \{x \in X_{(2)} \mid x_0 = 0\}$ and $[1]_0 = \{x \in X_{(2)} \mid x_0 = 1\}$ are cylinder sets. The partition $\sigma_{(2)}^{-k}(\xi)$ is independent of $\bigvee_{j=0}^{k-1} \sigma_{(2)}^{-j}(\xi)$ for all $k \geqslant 1$, so

$$h_{\mu_{(2)}}(\sigma_{(2)}, \xi) = \lim_{n\to\infty} \frac{1}{n} H_{\mu_{(2)}}\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right) = \lim_{n\to\infty} \frac{1}{n} \log 2^n = \log 2.$$

1.3.1 Elementary Properties

Notice that we are not yet in a position to compute $h_{\mu_{(2)}}(\sigma_{(2)})$ from Example 1.15, since this is defined as the supremum over all partitions in order to make the definition independent of the choice of ξ. In order to calculate $h_{\mu_{(2)}}(\sigma_{(2)})$ the basic properties of entropy need to be developed further.

Proposition 1.16. *Let $(X, \mathscr{B}, \mu, T)$ be a measure-preserving system on a Borel probability space, and let ξ and η be countable partitions of X with finite entropy. Then*

(1) $h_\mu(T, \xi) \leqslant H_\mu(\xi)$;
(2) $h_\mu(T, \xi \vee \eta) \leqslant h_\mu(T, \xi) + h_\mu(T, \eta)$;
(3) $h_\mu(T, \eta) \leqslant h_\mu(T, \xi) + H_\mu(\eta|\xi)$;
(4) $h_\mu(T, \xi) = h_\mu(T, \bigvee_{i=0}^{k} T^{-i}\xi)$ *for all* $k \geqslant 1$;
(5) *for invertible* T, $h_\mu(T, \xi) = h_\mu(T^{-1}, \xi) = h_\mu\left(T, \bigvee_{i=-k}^{k} T^{-i}\xi\right)$ *for all* $k \geqslant 1$;

Proof. In this proof we will make use of Proposition 1.7 without particular reference. These basic properties of entropy will be used repeatedly later.
(1): For any $n \geqslant 1$,

$$\begin{aligned}\frac{1}{n} H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right) &\leqslant \frac{1}{n}\sum_{i=0}^{n-1} H_\mu(T^{-i}\xi) \\ &= \frac{1}{n}\sum_{i=0}^{n-1} H_\mu(\xi) = H_\mu(\xi).\end{aligned}$$

(2): For any $n \geqslant 1$,

$$\begin{aligned}\frac{1}{n} H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}(\xi \vee \eta)\right) &= \frac{1}{n} H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right) + \frac{1}{n} H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\eta \Big| \bigvee_{i=0}^{n-1} T^{-i}\xi\right) \\ &\leqslant \frac{1}{n} H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right) + \frac{1}{n} H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\eta\right).\end{aligned}$$

(3): For any $n \geqslant 1$,

$$\begin{aligned}
h_\mu(T,\eta) &= \lim_{n\to\infty}\frac{1}{n}H_\mu\left(\bigvee_{i=0}^{n-1}T^{-i}\eta\right)\\
&\leqslant \lim_{n\to\infty}\frac{1}{n}H_\mu\left(\bigvee_{i=0}^{n-1}T^{-i}(\xi\vee\eta)\right)\\
&= h_\mu(T,\xi\vee\eta)\\
&= \lim_{n\to\infty}\left[\frac{1}{n}H_\mu\left(\bigvee_{i=0}^{n-1}T^{-i}\xi\right)+\frac{1}{n}H_\mu\left(\bigvee_{i=0}^{n-1}T^{-i}\eta\Big|\bigvee_{i=0}^{n-1}T^{-i}\xi\right)\right]\\
&\leqslant h_\mu(T,\xi)+\lim_{n\to\infty}\underbrace{\frac{1}{n}\sum_{i=0}^{n-1}H_\mu(T^{-i}\eta\big|T^{-i}\xi)}_{=H_\mu(\eta|\xi)}
\end{aligned}$$

by Lemma 1.12.
(4): For any $k\geqslant 1$,

$$\begin{aligned}
h_\mu(T,\bigvee_{i=0}^{k}T^{-i}\xi) &= \lim_{n\to\infty}\frac{1}{n}H_\mu\left(\bigvee_{j=0}^{n-1}T^{-j}\left(\bigvee_{i=0}^{k}T^{-i}\xi\right)\right)\\
&= \lim_{n\to\infty}\frac{1}{n}H_\mu\left(\bigvee_{i=0}^{k+n-1}T^{-i}\xi\right)\\
&= \lim_{n\to\infty}\left(\frac{k+n-1}{n}\right)\frac{1}{k+n-1}H_\mu\left(\bigvee_{i=0}^{k+n-1}T^{-i}\xi\right)=h_\mu(T,\xi).
\end{aligned}$$

(5): For any $n\geqslant 1$, Lemma 1.12 shows that

$$H_\mu\left(\bigvee_{i=0}^{n-1}T^{i}\xi\right)=H_\mu\left(T^{-(n-1)}\bigvee_{i=0}^{n-1}T^{i}\xi\right)=H_\mu\left(\bigvee_{i=0}^{n-1}T^{-i}\xi\right).$$

Dividing by n and taking the limit gives the first statement, and the second equality follows easily along the lines of (4). □

Proposition 1.17. *Let $(X,\mathscr{B},\mu,T)$ be a measure-preserving system on a Borel probability space. Then*

(1) $h_\mu(T^k)=kh_\mu(T)$ *for* $k\geqslant 1$.
(2) $h_\mu(T)=h_\mu(T^{-1})$ *if T is invertible.*

Proof.(1): For any partition ξ with finite entropy,

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}H_\mu\left(\bigvee_{j=0}^{n-1}T^{-kj}\left(\bigvee_{i=0}^{k-1}T^{-i}\xi\right)\right) &= \lim_{n\to\infty}\frac{k}{nk}H_\mu\left(\bigvee_{i=0}^{nk-1}T^{-i}\xi\right)\\
&= kh_\mu(T,\xi).
\end{aligned}$$

It follows that

$$h_\mu\left(T^k, \bigvee_{i=0}^{k-1} T^{-i}\xi\right) = kh_\mu(T,\xi),$$

so $kh_\mu(T) \leqslant h_\mu(T^k)$.

For the reverse inequality, notice that

$$h_\mu(T^k,\eta) \leqslant h_\mu(T^k, \bigvee_{i=0}^{k-1} T^{-i}\eta) = kh_\mu(T,\eta),$$

so $h_\mu(T^k) \leqslant kh_\mu(T)$.

(2): This follows from Proposition 1.16(5). □

Lemma 1.18. *Entropy can be computed using finite partitions only, in the sense that*

$$\sup_{\eta \text{ finite}} h_\mu(T,\eta) = \sup_{\xi: H_\mu(\xi)<\infty} h_\mu(T,\xi).$$

Proof. Any finite partition has finite entropy, so

$$\sup_{\eta \text{ finite}} h_\mu(T,\eta) \leqslant \sup_{\xi: H_\mu(\xi)<\infty} h_\mu(T,\xi).$$

For the reverse inequality, let ξ be any partition with $H_\mu(\xi) < \infty$. We claim that, for any $\varepsilon > 0$, we may find a finite partition η that is measurable with respect to $\sigma(\xi)$ and has $H_\mu\left(\xi\middle|\eta\right) < \varepsilon$. To see this, let $\xi = \{A_1, A_2, \dots\}$ and define

$$\eta = \{A_1, A_2, \dots, A_N, B_N = X \setminus \bigcup_{n=1}^{N} A_n\},$$

so that $\mu(B_N) \to 0$ as $N \to \infty$. Then

$$\begin{aligned} H_\mu(\xi\big|\eta) &= \mu(B_N) H_\mu\left(\frac{\mu(A_{N+1})}{\mu(B_N)}, \frac{\mu(A_{N+2})}{\mu(B_N)}, \dots\right) \\ &= -\sum_{j=N+1}^{\infty} \mu(A_j) \log \frac{\mu(A_j)}{\mu(B_N)} \\ &= -\sum_{j=N+1}^{\infty} \mu(A_j)\log\mu(A_j) + \underbrace{\mu(B_N)\log\mu(B_N)}_{\phi(B_N)}. \end{aligned}$$

Hence, by the assumption that $H_\mu(\xi) < \infty$ and by continuity of ϕ, it is possible to choose N large enough to ensure that $H_\mu(\xi\big|\eta) < \varepsilon$. It follows that

$$h_\mu(T,\xi) \leqslant h_\mu(T,\eta) + \varepsilon$$

by Proposition 1.16(3). □

1.3.2 Entropy as an Invariant

Recall that $(Y, \mathscr{B}_Y, \nu, S)$ is a *factor* of $(X, \mathscr{B}, \mu, T)$ if there is a measure-preserving map $\phi : X \to Y$ with $\phi(Tx) = S(\phi x)$ for μ-almost every $x \in X$.

Theorem 1.19. *If $(Y, \mathscr{B}_Y, \nu, S)$ is a factor of the system $(X, \mathscr{B}, \mu, T)$, then*

$$h_\nu(S) \leqslant h_\mu(T).$$

In particular, entropy is an invariant of measurable isomorphism.

Proof. Let $\phi : X \to Y$ be the factor map. Then any partition ξ of Y defines a partition $\phi^{-1}(\xi)$ of X, and since ϕ preserves the measure,

$$H_\nu(\xi) = H_\mu(\phi^{-1}(\xi)).$$

This immediately implies that $h_\mu(T, \phi^{-1}(\xi)) = h_\nu(S, \xi)$, and hence the result. □

The definition of the entropy of a measure-preserving transformation involves a supremum over the set of all (finite) partitions. In order to compute the entropy, it is easier to work with a single partition. The next result – the Kolmogorov–Sinaĭ Theorem – gives a sufficient condition on a partition to allow this.

Theorem 1.20. *Suppose $(X, \mathscr{B}, \mu, T)$ is a measure-preserving system on a Borel probability space, and ξ is a partition of finite entropy that is a* one-sided generator *under T in the sense that*

$$\bigvee_{n=0}^{\infty} T^{-n}\xi = \mathscr{B}. \tag{1.8}$$

Then

$$h_\mu(T) = h_\mu(T, \xi).$$

If T is invertible and ξ is a partition with finite entropy that is a generator *under T in the sense that*

$$\bigvee_{n=-\infty}^{\infty} T^{-n}\xi = \mathscr{B}.$$

Then

$$h_\mu(T) = h_\mu(T, \xi).$$

Lemma 1.21. *Let $(X, \mathscr{B}, \mu, T)$ and ξ be as in Theorem* 1.20*, and let η be another finite partition of X. Then*

$$H_\mu\left(\eta \middle| \bigvee_{i=0}^{n} T^{-i}\xi\right) \to 0$$

as $n \to \infty$.

PROOF. By assumption, the partitions $\bigvee_{i=0}^{n} T^{-i}\xi$ for $n = 1, 2, \ldots$ together generate $\mathscr{B}$. This shows that for any $\delta > 0$ and $B \in \mathscr{B}$, there exists some n and some set $A \in \bigvee_{i=0}^{n} T^{-i}\xi$ for which $\mu(A \triangle B) < \delta$. Applying this to all the elements of $\eta = \{B_1, \ldots, B_m\}$ we can find one n with the property that $\mu(A_i' \triangle B_i) < \delta/m$ for $i = 1, \ldots, m$ and some $A_i' \in \bigvee_{i=0}^{n} T^{-i}\xi$. Write

$$A_i = A_1', A_2 = A_2' \setminus A_1, \ldots, A_m = A_m' \setminus \bigcup_{j=1}^{m} A_j',$$

and notice that

$$\begin{aligned}\mu(A_i \triangle B_i) &= \mu(A_i \setminus B_i) + \mu(B_i \setminus A_i) \\ &\leqslant \mu(A_i' \setminus B_i) + \mu(B_i \setminus A_i) + \mu(B_i \cap \bigcup_{j=1}^{i-1} A_j') \\ &\leqslant \frac{\delta}{m} + \sum_{j=1}^{i-1} \mu(A_j' \setminus B_j) \leqslant \delta\end{aligned}$$

by construction. Let $\zeta = \{A_1, \ldots, A_m\}$, so that

$$\begin{aligned}H_\mu\left(\eta \middle| \bigvee_{i=1}^{n} T^{-1}\xi\right) &\leqslant H_\mu\left(\eta \middle| \zeta\right) \quad \text{(by Prop. 1.7(4))} \\ &= -\sum_{i=1}^{m} \mu(A_i \cap B_i) \log \frac{\mu(A_i \cap B_i)}{\mu(A_i)} \\ &\qquad - \sum_{i,j=1, i\neq j}^{m} \mu(A_i \cap B_j) \log \frac{\mu(A_i \cap B_j)}{\mu(A_i)}.\end{aligned}$$

The terms in the first sum are close to zero because $\frac{\mu(A_i \cap B_i)}{\mu(A_i)}$ is close to 1, and the terms in the second sum are close to zero because $\mu(A_i \cap B_j)$ is close to zero. In other words, given any $\varepsilon > 0$, by choosing δ small enough (and hence n large enough) we can ensure that

$$H_\mu\left(\eta \middle| \bigvee_{i=0}^{n} T^{-i}\xi\right) < \varepsilon$$

as needed. □

PROOF OF THEOREM 1.20. Let ξ be a one-sided generator under T. For any partition η,

$$h_\mu(T, \eta) \leqslant \underbrace{h_\mu(T, \bigvee_{i=0}^{n} T^{-i}\xi)}_{=h_\mu(T,\xi)} + \underbrace{H_\mu\left(\eta \middle| \bigvee_{i=0}^{n} T^{-i}\xi\right)}_{\to 0 \text{ as } n\to\infty}$$

by Proposition 1.16(3) and Lemma 1.21, so $h(T,\eta) \leqslant h(T,\xi)$ as required. The proof for a generator under an invertible T is similar. □

Corollary 1.22. *If $(X,\mathscr{B},\mu,T)$ is an invertible measure-preserving system on a Borel probability space with a one-sided generator, then $h_\mu(T)=0$.*

Proof. Let ξ be a partition with

$$\bigvee_{n=0}^{\infty} T^{-n}\sigma(\xi) \underset{\mu}{=} \mathscr{B},$$

so that

$$h_\mu(T) = h_\mu(T,\xi) = H_\mu\left(\xi \Big| \bigvee_{i=1}^{\infty} T^{-i}\xi\right)$$

by Theorem 1.20 and Proposition 1.16(4). On the other hand, since T is invertible we have

$$\bigvee_{i=1}^{\infty} T^{-i}\xi = T^{-1}\left(\bigvee_{i=0}^{\infty} T^{-i}\xi\right) \underset{\mu}{=} T^{-1}\mathscr{B} = \mathscr{B},$$

(the hypothesis of invertibility is used in the last equality) so

$$h_\mu(T) = H_\mu\left(\xi \Big| \bigvee_{i=1}^{\infty} T^{-i}\xi\right) = 0$$

by Lemma 1.21. □

The Kolmogorov–Sinaĭ theorem allows the entropy of simple examples to be computed. The next examples will indicate how positive entropy arises, and gives some indication that the entropy of a transformation is related to the complexity of its orbits. In Examples 1.25 and 1.26 the positive entropy reflects the way in which the transformation moves nearby points apart and thereby chops up the space in a complicated way; in Examples 1.23 and 1.24 the transformation moves points around in a very orderly way, and this is reflected[(5)] in the zero entropy.

Example 1.23. The identity map $I : X \to X$ has zero entropy on any probability space $(X,\mathscr{B},\mu)$. This is clear, since for any partition ξ, $\bigvee_{i=0}^{n-1} I^{-i}\xi = \xi$, so $h_\mu(I,\xi)=0$.

Example 1.24. The circle rotation $R_\alpha : \mathbb{T} \to \mathbb{T}$ has zero entropy with respect to Lebesgue measure. If α is rational, then there is some $q \geqslant 1$ with $R_\alpha^q = I$, so $h_{m_{\mathbb{T}}}(R_\alpha) = 0$ by Proposition 1.17(1) and Example 1.23. If α is irrational, then $\xi = \{[0,\frac{1}{2}),[\frac{1}{2},1)\}$ is a one-sided generator since the point 0 has dense orbit under R_α. In fact, if $x_1, x_2 \in \mathbb{T}$ with $x_1 < x_2 \in [0,\frac{1}{2})$ as real numbers, then there is some $n \in \mathbb{N}$ with $R_\alpha^n(0) \in (x_1,x_2)$, or equivalently $x_2 \in R_\alpha^{-n}[0,\frac{1}{2})$ but $x_1 \in R_\alpha^{-n}[\frac{1}{2},1)$. This implies that the atoms of $\bigvee_{n=0}^{\infty} T^{-n}\xi$ consist of single points. It follows that $h_{m_{\mathbb{T}}}(R_\alpha) = 0$ by Corollary 1.22.

Example 1.25. The state partition for the Bernoulli 2-shift in Example 1.15 is a two-sided generator, so we deduce that $h_{\mu_2}(\sigma_2) = \log 2$.

The state partition

$$\big\{\{x \in X_{(3)} \mid x_0 = 0\}, \{x \in X_{(3)} \mid x_0 = 1\}, \{x \in X_{(3)} \mid x_0 = 2\}\big\}$$

of the Bernoulli 3-shift $X_{(3)} = \{0,1,2\}^{\mathbb{Z}}$ with the $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ measure $\mu_{(3)}$ is a two-sided generator under the left shift $\sigma_{(3)}$, so the same argument shows that $h_{\mu_{(3)}}(\sigma_{(3)}) = \log 3$. Thus the Bernoulli 2- and 3-shifts are not measurably isomorphic.

Example 1.26. The partition $\xi = \{[0, \frac{1}{2}), [\frac{1}{2}, 1)$ is a one-sided generator for the circle-doubling map $T_2 : \mathbb{T} \to \mathbb{T}$. It is easy to check that $\bigvee_{i=0}^{n-1} T^{-i}\xi$ is the partition

$$\{[0, \tfrac{1}{2^n}), \ldots, [\tfrac{2^n-1}{2^n}, 1)\},$$

so $H_{m_{\mathbb{T}}}\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right) = \log 2^n$. Theorem 1.20 shows that $h_{m_{\mathbb{T}}}(T_2) = \log 2$.

Example 1.27. Just as in Example 1.26, the partition

$$\xi = \{[0, \tfrac{1}{p}), [\tfrac{1}{p}, \tfrac{2}{p}), \ldots, [\tfrac{p-1}{p}, 1)\}$$

is a generator for the map $T_p(x) = px \pmod 1$ with $p \geqslant 2$ on the circle, and a similar argument shows that $h_{m_{\mathbb{T}}}(T_p) = \log p$.

Now consider an arbitrary T_p-invariant probability measure μ on the circle. Since ξ is a generator (indeed, equation (1.8) holds without reference to a measure), we have

$$h_\mu(T_p) = h_\mu(T_p, \xi) \leqslant H_\mu(\xi) \leqslant \log p \tag{1.9}$$

by Proposition 1.16(1) and Proposition 1.5, since ξ has only p elements.

Let us now *characterize* those measures for which we have equality in the estimate equation (1.9). By Lemma 1.13,

$$h_\mu(T_p, \xi) = \inf_{n \geqslant 1} \frac{1}{n} H_\mu\left(\xi \vee T_p^{-1}\xi \vee \cdots \vee T_p^{-(n-1)}\xi\right) \leqslant \frac{1}{n} \log p^n,$$

where the last inequality holds by Proposition 1.5. Hence, $h_\mu(T_p) = \log p$ implies, using the equality case in Proposition 1.5, that each of the intervals $[\frac{j}{p^n}, \frac{j+1}{p^n})$ of the partition $\xi \vee T_p^{-1}\xi \vee \cdots \vee T_p^{-(n-1)}\xi$ must have μ-measure equal to $\frac{1}{p^n}$. This implies that $\mu = m_{\mathbb{T}}$, thus characterizing $m_{\mathbb{T}}$ as the only T_p-invariant Borel probability measure with entropy equal to $\log p$.

The phenomenon seen in Example 1.27, where maximality of entropy can be used to characterize particular measures is important, and it holds in other situations too. In this case, the geometry of the generating partition is very simple. In other contexts, it is often impossible to pick a generator that is so

convenient. Apart from these complications arising from the geometry of the space and the transformation, the phenomenon that maximality of entropy can be used to characterize certain measures always goes back to the strict convexity of the map $x \mapsto -x \log x$. We will see other instances of this in Section 4.5, and in Chapter 4 we will further develop the machinery of entropy. In particular, we will study the relationship between entropy and ergodic decomposition.

Exercises for Section 1.3

Exercise 1.3.1. For a sequence of finite partitions (ξ_n) with $\sigma(\xi_n) \nearrow \mathscr{B}$, prove that $h(T) = \lim_{n\to\infty} h(T, \xi_n)$.

Exercise 1.3.2. Prove that $h_{\mu\times\nu}(T \times S) = h_\mu(T) + h_\nu(S)$.

Exercise 1.3.3. Show that there exists a shift-invariant measure μ on the shift space $X = \{0,1\}^{\mathbb{Z}}$ with $h_\mu(\sigma) = 0$ and with full support on X.

Exercise 1.3.4. Let $(X, \mathscr{B}, \mu, T)$ be a measure-preserving system, and let ξ be a countable partition of X with finite entropy. Show that $\frac{1}{n} H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right)$ decreases to $h(T, \xi)$ by the following steps.

(1) Use Proposition 1.7(1) to show that

$$H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right) = H_\mu(\xi) + \sum_{j=1}^{n-1} H_\mu\left(\xi \Big| \bigvee_{i=1}^{j} T^{-i}\xi\right)$$

by induction.

(2) Deduce that

$$H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right) \geqslant n H_\mu\left(\xi \Big| \bigvee_{i=1}^{n} T^{-i}\xi\right)$$

(3) Use (2) and Proposition 1.7(1) again to show that

$$n H_\mu\left(\bigvee_{i=0}^{n} T^{-i}\xi\right) \leqslant (n+1) H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right)$$

and deduce the result.

Exercise 1.3.5. Show that $h_\mu(T, \xi)$ is a continuous function of ξ in the L^1_μ norm on ξ.

Exercise 1.3.6. (a) Show that the entropy of the Bernoulli shift defined by the probability vector $\mathbf{p} = (p_1, \ldots, p_n)$ is given by $-\sum_{i=1}^n p_i \log p_i$ (see [38, Ex. 2.9]).
(b) Deduce that, for any $h \in [0, \infty)$, there is a measure-preserving transformation with entropy h.
(c) Show that the entropy is maximized when $\mathbf{p} = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$.

1.4 Defining Entropy using Names

We saw in Section 1.1 that the entropy formula in Definition 1.1 is the unique formula satisfying the basic properties of information from Section 1.1.2. In this section we describe another way in which Definition 1.1 is forced on us, by computing a quantity related to entropy for a Bernoulli shift. For a measure-preserving system $(X, \mathscr{B}, \mu, T)$ and a partition $\xi = (A_1, A_2, \ldots)$ (thought of as an ordered list), define the (ξ, n)-name $\mathbf{w}_n^\xi(x)$ of a point $x \in X$ to be the vector $(a_0, a_1, \ldots, a_{n-1})$ with the property that $T^i(x) \in A_{a_i}$ for $0 \leqslant i < n$. We also denote by $\mathbf{w}_n^\xi(x)$ the set of all points that share the (ξ, n)-name of x, which is clearly the atom of x with respect to $\bigvee_{i=0}^{n-1} T^{-i}\xi$. By definition, the entropy of a measure-preserving transformation is related to the distribution of the measures of the names. We claim that this relationship goes deeper, by showing that the logarithmic rate of decay of the volume of a typical name is the entropy. This is the content of the Shannon–McMillan–Breiman theorem (Theorem 6.1).

In this section we compute the rate of decay of the measure of names for a Bernoulli shift, which will serve both as another motivation for Definition 1.1 and as a forerunner of Theorem 6.1.

Lemma 1.28. *Let $(X, \mathscr{B}, \mu, T)$ be the Bernoulli shift defined by the probability vector $(p_1, \ldots, p_s)$, so that $X = \prod_{\mathbb{Z}}\{1, \ldots, s\}$, $\mu = \prod_{\mathbb{Z}}(p_1, \ldots, p_s)$, and T is the left shift. Then*

$$\frac{1}{n}\log\mu\left(\mathbf{w}_n^\xi(x)\right) \to -\sum_{i=1}^{s} p_i \log p_i$$

as $n \to \infty$ for μ-almost every x.

Proof. The set of points with the name $\mathbf{w}_n^\xi(x)$ is the cylinder set

$$\{y \in X \mid y_0 = x_0, \ldots, y_{n-1} = x_{n-1}\},$$

so

$$\mu\left(\mathbf{w}_n^\xi(x)\right) = p_{x_0} \cdots p_{x_{n-1}}. \tag{1.10}$$

Now for $1 \leqslant j \leqslant s$, write $\chi_j = \chi_{[j]_0}$ (where $[j]_0$ denotes the cylinder set of points with 0 coordinate equal to j) and notice that

$$\sum_{i=0}^{n-1} \chi_j(T^i x) = \left|\{i \mid 0 \leqslant i \leqslant n-1, x_i = j\}\right|,$$

so we may rearrange equation (1.10) to obtain

$$\mu\left(\mathbf{w}_n^\xi(x)\right) = p_1^{\sum_{i=0}^{n-1}\chi_1(T^i x)} p_2^{\sum_{i=0}^{n-1}\chi_2(T^i x)} \cdots p_s^{\sum_{i=0}^{n-1}\chi_s(T^i x)}. \tag{1.11}$$

Now, by the ergodic theorem, for any $\varepsilon > 0$ and for almost every $x \in X$ there is an N so that for every $n \geqslant N$ we have

$$\left| \frac{1}{n} \sum_{i=0}^{n-1} \chi_j(T^i x) - p_j \right| < \varepsilon.$$

Thus in equation (1.11) we already see – to within a small error – the expression $(p_1^{p_1} \cdots p_s^{p_s})^n$, whose logarithmic decay rate will give us the familiar entropy formula in Definition 1.1. By equation (1.11) we deduce that

$$\left| \log \mu \left(\mathbf{w}_n^\xi(x) \right) - n \log(p_1^{p_1} \cdots p_s^{p_s}) \right| \leqslant \varepsilon n |\log(p_1 \cdots p_s)|. \tag{1.12}$$

Now given $\varepsilon' > 0$, choose $\varepsilon > 0$ small enough to ensure that

$$\varepsilon |\log(p_1 \cdots p_s)| < \varepsilon'$$

and then equation (1.12) shows that

$$\left| \frac{1}{n} \log \mu \left(\mathbf{w}_n^\xi(x) \right) - \log(p_1^{p_1} \cdots p_s^{p_s}) \right| \leqslant \varepsilon |\log(p_1 \cdots p_s)| < \varepsilon'$$

almost everywhere, so

$$\frac{1}{n} \log \mu \left(\mathbf{w}_n^\xi(x) \right) \to - \sum_{i=1}^{s} p_i \log p_i$$

as $n \to \infty$. □

1.4.1 Name Entropy

In fact the entropy theory for measure-preserving transformations can be built up entirely in terms of names, and this is done in the elegant monograph by Rudolph [125, Chap. 5]. We only discuss this approach briefly, and will not use the following discussion in the remainder of the book (entropy is such a fecund notion that similar alternative entropy notions will arise several times: see the definition of topological entropy using open covers in Section 2.2, Theorem 6.1, and Section 3.3).

Let $(X, \mathscr{B}, \mu, T)$ be an ergodic* measure-preserving transformation, and define for each finite partition $\xi = \{A_1, \ldots, A_r\}$, $\varepsilon > 0$ and $n \geqslant 1$ a quantity $N(\xi, \varepsilon, n)$ as follows. For each (ξ, n)-name $\mathbf{w}^\xi \in \{1, \ldots, r\}^n$ write $\mu(\mathbf{w}^\xi)$ for the measure $\mu\left(\{x \in X \mid \mathbf{w}_n^\xi(x) = \mathbf{w}^\xi\}\right)$ of the set of points in X whose name is $\mathbf{w}^\xi$, where $\mathbf{w}_n^\xi(x) = (a_0, \ldots, a_{n-1})$ with $T^j(x) \in A_{a_j}$ for $0 \leqslant j < n$. Starting with the names of least measure in $\{1, \ldots, r\}^n$, remove as many

* To obtain an independent and equivalent definition in the way described here, ergodicity needs to be assumed initially.

names as possible compatible with the condition that the total measure of the remaining names exceeds $(1-\varepsilon)$. Write $N(\xi,\varepsilon,n)$ for the cardinality of the set of remaining names. Then one may define

$$h_{\mu,\mathrm{name}}(T,\xi)=\lim_{\varepsilon\to 0}\liminf_{n\to\infty}\frac{1}{n}\log N(\xi,\varepsilon,n)$$

and

$$h_{\mu,\mathrm{name}}(T)=\sup_{\xi}h_{\mu,\mathrm{name}}(T,\xi)$$

where the supremum is taken over all finite partitions. Using this definition and the assumption of ergodicity, it is possible to prove directly the following basic theorems:

(1) The Shannon–McMillan–Breiman theorem (Theorem 6.1) in the form

$$-\frac{1}{n}\log\mu\left(\mathbf{w}_n^{\xi}(x)\right)\longrightarrow h_{\mu,\mathrm{name}}(T,\xi) \tag{1.13}$$

for μ-almost every x.

(2) The Kolmogorov–Sinaĭ theorem: if $\bigvee_{i=-\infty}^{\infty}T^{-i}\xi=\mathscr{B}$, then

$$h_{\mu,\mathrm{name}}(T,\xi)=h_{\mu,\mathrm{name}}(T). \tag{1.14}$$

We shall see later that equation (1.13) shows $h_{\mu,\mathrm{name}}(T,\xi)=h_{\mu}(T,\xi)$, by Theorem 6.1.

In contrast to the development in Sections 1.1–1.3, the formula in Definition 1.1 is not used in defining $h_{\mu,\mathrm{name}}$. Instead it appears as a consequence of the combinatorics of counting names as in Lemma 1.28.

1.5 Compression Rate

Recall from Section 1.2 the interpretation of the entropy $\frac{1}{\log 2}H_{\mu}(\xi)$ as the optimal average length of binary codes compressing the possible outcomes of the experiment represented by the partition ξ (ignoring the failure of optimality by one digit, as in Lemma 1.11).

This interpretation also helps to interpret some of the results of this section. For example, the subadditivity

$$H_{\mu}\left(\bigvee_{i=0}^{n-1}T^{-i}\xi\right)\leqslant nH_{\mu}(\xi)$$

can be interpreted to mean that the almost optimal code as in Lemma 1.11 for $\xi=(A_1,A_2,\dots)$ can be used to code $\bigvee_{i=0}^{n-1}T^{-i}\xi$ as follows. The partition $\bigvee_{i=0}^{n-1}T^{-i}\xi$ has as a natural alphabet the names $i_0\dots i_{n-1}$ of length n in the alphabet of ξ. The requirements on codes ensures that the code induces in a natural way a code s_n on names of length n by concatenation,

$$s_n(i_0 \dots i_{n-1}) = s(i_0)s(i_1)\dots s(i_{n-1}). \tag{1.15}$$

The average length of this code is $nH_\mu(\xi)$. However (unless the partitions $\xi, T^{-1}\xi, \dots, T^{-(n-1)}\xi$ are independent), there might be better codes for names of length n than the code s_n constructed by equation (1.15), giving the subadditivity inequality by Lemma 1.10.

Thus

$$\frac{1}{n} H_\mu \left(\bigvee_{i=0}^{n-1} T^{-i}\xi \right)$$

is the average length of the optimal code for $\bigvee_{i=0}^{n-1} T^{-i}\xi$ averaged both over the space and over a time interval of length n. Moreover, $h_\mu(T, \xi)$ is the lowest averaged length of the code per time unit describing the outcomes of the experiment ξ on long pieces of trajectories that could possibly be achieved. Since $h_\mu(T, \xi)$ is defined as an infimum in Definition 1.14, this might not be attained, but any slightly worse compression rate would be attainable by working with sufficiently long blocks $T^{-km} \bigvee_{i=0}^{m-1} T^{-i}\xi$ of a much longer trajectory in $\bigvee_{i=0}^{nm-1} T^{-i}\xi$. Notice that the slight lack of optimality in Lemmas 1.10 and 1.11 vanishes on average over long time intervals (see Exercise 1.5.3).

Example 1.29. Consider the full three shift $\sigma_{(3)} : \{0,1,2\}^{\mathbb{Z}} \to \{0,1,2\}^{\mathbb{Z}}$, with the generator $\xi = \{[0]_0, [1]_0, [2]_0\}$ (using the notation from Exercise 1.15 for cylinder sets). A code for ξ is

$$\begin{aligned} 0 &\longmapsto 00, \\ 1 &\longmapsto 01, \\ 2 &\longmapsto 10, \end{aligned}$$

which gives a rather inefficient coding for names: the length of a ternary sequence encoded in this way doubles. Using blocks of ternary sequences of length 3 (with a total of 27 sequences) gives binary codes of length 5 (out of a total 32 possible codes), showing the greater efficiency in longer blocks. Defining a code by some injective map $\{0,1,2\}^3 \to \{0,1\}^5$ allows a ternary sequence of length $3k$ to be encoded to a binary sequence of length $5k$, giving the better ratio of $\frac{5}{3}$. Clearly these simple codes will never give a better ratio that $\frac{\log 3}{\log 2}$, but can achieve any slightly larger ratio at the expense of working with very long blocks of sequences.

One might wonder whether more sophisticated codes could, on average, be more efficient on long sequences. The results of this chapter say precisely that this is not possible if we assume that the digits in the ternary sequences considered are identically independently distributed; equivalently if we work with the system $(X_{(3)}, \mu_3, \sigma_{(3)})$ with entropy $h_{\mu_3}(\sigma_{(3)}) = \log 3$.

Exercises for Section 6.2

Exercise 1.5.1. Give an interpretation of the finiteness of the entropy of an infinite probability vector $(v_1, v_2, \dots)$ in terms of codes.

Exercise 1.5.2. Give an interpretation of conditional entropy and information in terms of codes.

Exercise 1.5.3. Fix a finite partition ξ with corresponding alphabet A in an ergodic measure-preserving system $(X, \mathscr{B}, \mu, T)$, and for each $n \geqslant 1$ let s_n be an optimal prefix code for the blocks of length n over A. Use the source coding theorem to show that

$$\lim_{n\to\infty} \frac{\log 2}{n} L(s_n) = h(T, \xi).$$

1.6 Entropy and Classification

For our purposes, entropy will be used primarily as a tool to understand properties of measures in a dynamical system. However, the original motivation for defining entropy comes about through its invariance properties and its role in determining the structure of certain kinds of measure-preserving systems. The most important part of this theory is due to Ornstein, and in this section we give a short introduction to this[6]. We will not be using the results in this section, so proofs and even exact statements are omitted. In this section partitions are to be thought of as ordered lists of sets. Before describing this, we mention a simple example of a family of isomorphisms found by Mešalkin.

Example 1.30. As mentioned on p. 3, Mešalkin [96] found some special cases of isomorphisms between Bernoulli shifts. Let $\mathsf{X} = (X, \mathscr{B}, \mu, \sigma_1)$ be the Bernoulli shift with a state space of 4 symbols and measure $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$; let $\mathsf{Y} = (Y, \mathscr{C}, \nu, \sigma_2)$ be the Bernoulli shift with a state space of 5 symbols and measure $(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$. Notice that the state partition is a generator, so just as in Example 1.25 we can show that

$$h_\mu(\mathsf{X}) = h_\nu(\mathsf{Y}) = \log 4.$$

Mešalkin showed[7] that X and Y are isomorphic, by constructing an invertible measure-preserving map $\phi : X \to Y$ with $\phi\sigma_1 = \sigma_2\phi$ μ-almost everywhere. The following way of understanding Mešalkin's isomorphism is due to Jakobs [64] and we learnt it from Benjamin Weiss. Write the alphabet of the Bernoulli shift X as

$$\begin{array}{c} 0\ \ 1\ \ 0\ \ 1 \\ 0,\ 0,\ 1,\ 1. \end{array}$$

For the shift Y, use the alphabet

$$\begin{array}{c} 0\ 0\ 1\ 1 \\ 0\ 1\ 0\ 1 \\ 0,\,1,\,1,\,1,\,1, \end{array}$$

with measures $\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}$ respectively. A typical point $y = (y_n) \in Y$ is shown in Figure 1.3. View the short blocks 0 as poor people, and the tall blocks as wealthy ones.

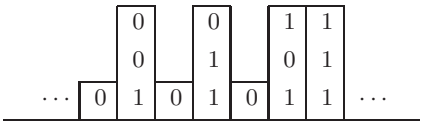

Fig. 1.3. A typical point in the $(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$ Bernoulli shift.

The shift X is egalitarian: all symbols have equal height. Construct a map from Y to X by requiring that each wealthy person in y find a poor neighbor and give her or him a symbol according to the following procedure.

- If a wealthy person has a poor neighbor immediately to her or his right, the person donates the top symbol to that neighbor, for example:

$$\begin{array}{l} 0 \\ 1 \\ 1\ 0 \end{array} \quad \longrightarrow \quad 1\ 0 \qquad 1\ 0$$

- If the neighbor to the immediate right is wealthy too, the donation goes to the first poor person on the right who has not received a donation from a closer wealthy person in between them. In other words, in a poor neighborhood, like $\ldots 000 \ldots$, one needs to look left in the sequence y until a wealthy person is found who has not donated a symbol, and take the top symbol from her or him.

Elementary properties of the simple random walk (specifically, recurrence of the one-dimensional random walk; see for example Spitzer [135]) says that with probability one each poor person finds exactly one wealthy person to pair up with. This is the key step in proving that the map is an invertible measurable isomorphism. The inverse map redistributes wealth from the poor to the wealthy – this uses the fact that after the original redistribution of wealth one can still reconstruct who had been wealthy and who had been poor by using the bottom symbol.

Ornstein developed a way of studying partitions for measure-preserving systems that allowed him to determine when an abstract measure-preserving system is isomorphic to a Bernoulli shift, and decide when two Bernoulli shifts are isomorphic. In order to describe this theory, we start by saying a

little more about names. Let $(X, \mathscr{B}, \mu, T)$ be an invertible ergodic measure-preserving system on a Borel probability space, and fix a finite measurable partition

$$\xi = (A_1, \ldots, A_r)\,.$$

The partition ξ defines a map

$$\mathbf{w}^\xi : X \to Y = \{1, \ldots, r\}^{\mathbb{Z}}$$

by requiring that $(\mathbf{w}^\xi(x))_k = j$ if and only if $T^k x \in A_j$ for $k \in \mathbb{Z}$. Thus $\mathbf{w}^\xi(x)$ restricted to the coordinates $[0, n-1]$ is the usual (ξ, n)-name $\mathbf{w}_n^\xi(x)$. Clearly

$$\mathbf{w}^\xi(Tx) = \sigma(\mathbf{w}^\xi x),$$

where σ as usual denotes the left shift on Y. Write $\mathscr{B}_Y$ for the Borel σ-algebra (with the discrete topology on the alphabet $\{1, \ldots r\}$ and the product topology on Y), and define a measure ν on Y to be the push-forward of μ, so

$$\nu(A) = \mu\left((\mathbf{w}^\xi)^{-1}(A)\right)$$

for all $A \in \mathscr{B}_Y$. Thus

$$\mathbf{w}^\xi : \mathsf{X} = (X, \mathscr{B}, \mu, T) \to \mathsf{Y} = (Y, \mathscr{B}_Y, \nu, \sigma).$$

is a *factor map*. It is easy to show that $\mathbf{w}^\xi$ is an *isomorphism* if and only if ξ is a generator.

Definition 1.31. *A partition $\xi = \{A_1, \ldots, A_r\}$ is* independent *under T if for any choice of distinct $j_1, \ldots, j_k \in \mathbb{Z}$ and sets $A_{i_1}, \ldots, A_{i_k}$ we have*

$$\mu\left(T^{-j_1}A_{i_1} \cap T^{-j_2}A_{i_2} \cap \cdots \cap T^{-j_k}A_{i_k}\right) = \mu(A_{i_1})\mu(A_{i_2}) \cdots \mu(A_{i_k}).$$

Example 1.32. The state partition $\xi = \{[1]_0, [2]_0, \ldots, [r]_0\}$ in the Bernoulli shift $\{1, \ldots, r\}^{\mathbb{Z}}$ with shift-invariant measure $\mu = \prod_{i\in\mathbb{Z}}(p_1, \ldots, p_r)$ is independent under the shift.

Lemma 1.33. *An invertible measure-preserving system on a Borel probability space is isomorphic to a Bernoulli shift if and only if it has an independent generator.*

Notice that if ξ is an independent generator for $(X, \mathscr{B}, \mu, T)$ then

$$\begin{aligned} h_\mu(T) &= h_\mu(T, \xi) \qquad \text{(since } \xi \text{ is a generator)} \\ &= H_\mu(\xi) \qquad \text{(since } \xi \text{ is independent)}. \end{aligned}$$

Measure-preserving systems X and Y are said to be *weakly isomorphic* if each is a factor of the other. Theorem 1.19 really shows that entropy is an invariant of weak isomorphism. It is far from obvious, but true[(8)], that systems can be weakly isomorphic without being isomorphic. Sinaĭ showed [133] that

weakly isomorphic systems have the same entropy, are spectrally isomorphic, are isomorphic if they have discrete spectrum, and gave several other properties that they must share. He also proved in his paper [132] the deep result that if X is a Bernoulli shift and Y any ergodic system with $h(\mathsf{Y}) \geqslant h(\mathsf{X})$, then X is isomorphic to a factor of Y. Thus, for example, Bernoulli shifts of the same entropy are weakly isomorphic. Ornstein's isomorphism theorem (proved in [101] for finite entropy and extended to the infinite entropy case in [102]) strengthens this enormously by showing that Bernoulli shifts of the same entropy must be isomorphic.

Theorem (Ornstein). If $\mathsf{X} = (X, \mathscr{B}_X, \mu, T)$ and $\mathsf{Y} = (Y, \mathscr{B}_Y, \nu, S)$ are Bernoulli shifts, then X is isomorphic to Y if and only if $h_\mu(T) = h_\nu(S)$.

In general it seems very difficult to decide if a given system has an independent generator, so it is not clear how widely applicable the isomorphism theory is. One aspect of Ornstein's work is a series of strengthenings of Lemma 1.33 that make the property of being isomorphic to a Bernoulli shift something that can be checked, allowing a large class of measure-preserving systems to be shown to be isomorphic to Bernoulli shifts, and a series of results showing that the property of being isomorphic to a Bernoulli shift is preserved by taking factors or limits. Using the stronger characterizations of the property of being isomorphic to a Bernoulli shift, many important measure-preserving transformations are known to have this property (and are therefore measurably classified by their entropy). The next example describes some of these (and some simple examples that cannot be isomorphic to a Bernoulli shift). For brevity we will say a system "is a Bernoulli automorphism" to mean that it is isomorphic to a Bernoulli shift.

Example 1.34. (1) The automorphism of $\mathbb{T}^2$ associated to the matrix $\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$ (see Section 4.5) is a Bernoulli automorphism.

(2) More generally, Katznelson [71] showed that any ergodic toral automorphism is a Bernoulli automorphism. One of the critical estimates used in this argument has been simplified by Lind and Schmidt [86] using the product formula for global fields.

(3) More generally still, any ergodic automorphism of a compact group is a Bernoulli automorphism. This was proved independently by Lind [83] and Miles and Thomas [97], [99], [98]. Some simplifications were made by Aoki [6].

(4) A mixing Markov shift is a Bernoulli automorphism (see Ornstein and Shields [105].)

(5) Certain ergodic automorphisms of nilmanifolds are Bernoulli automorphisms (see Dani [28]).

(6) The map of geodesic flow for a fixed time on a surface of negative curvature is a Bernoulli automorphism (see [107]).

(7) The map defined by the flow for a fixed time of one billiard ball moving on a square table with finitely many convex obstacles is a Bernoulli automorphism (see Ornstein and Gallavotti [47]).
(8) The flow defined by the motion of hard elastic spheres in a box induces a Bernoulli automorphism (this result is due to Sinai; see Ornstein [104]).
(9) A generalization of (5) is that any mixing Anosov flow preserving a smooth measure is a Bernoulli automorphism (see Ratner [119] or Bunimovič [19]).
(10) Define a transformation $T_\beta : [0,1] \to [0,1]$ by $T_\beta(x) = \beta x \pmod 1$, where $\beta > 1$. This is the β-transformation. For each $\beta > 1$, there is a T_β-invariant measure on $[0,1]$ that is absolutely continuous with respect to Lebesgue measure, discovered by Rényi [121]. Let $\widetilde{T}_\beta$ denote the natural extension of T_β to an invertible measure-preserving transformation. Then $\widetilde{T}_\beta$ is a Bernoulli automorphism (see Smorodinsky [134] or Fischer [44]).
(11) Notice that a Bernoulli automorphism automatically has positive entropy (we exclude the map on a single point). It follows that a circle rotation, for example, is never a Bernoulli automorphism.

The definitive nature of Theorem 1.6 should not mask the scale of the problem of classifying measure-preserving transformations up to isomorphism: Bernoulli shifts are a significant class, encompassing many geometrically natural maps, but the structure of most measure-preserving systems remains mysterious[(9)].

Notes to Chapter 1

[(1)] (Page 3) The original material may be found in papers of Kolmogorov [77] (corrected in [76]), Rokhlin [122], and Rokhlin and Sinaĭ [123]. For an attractive survey of the foundations and later history of entropy in ergodic theory, see the survey article by Katok [70]. The concept of entropy is due originally to the physicist Clausius [25], who used it in connection with the dispersal of usable energy in thermodynamics in 1854 and coined the term 'entropy' in 1868. Boltzmann [11] later developed a statistical notion of entropy for ideal gases, and von Neumann a notion of entropy for quantum statistical mechanics; it remains an important concept in thermodynamics and statistical mechanics. The more direct precursor to the ergodic-theoretic notion of entropy comes from the work of Shannon [129] in information theory. The name 'entropy' for Shannon's measure of information carrying capacity was apparently suggested by von Neumann: Shannon is quoted by Tribus and McIrvine [138] as recalling that

> "My greatest concern was what to call it. I thought of calling it information, but the word was overly used, so I decided to call it uncertainty. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important,

nobody knows what entropy really is, so in a debate you will always have the advantage.' "

[(2)](Page 12) The connections between information theory and ergodic theory, many of which originate with Shannon [129], are pervasive. Accessible accounts may be found in the monographs of Choe [23], Shields [130], and Weiss [146]; particularly relevant results in two papers of Orstein and Weiss [108], [109]. Some aspects of coding and symbolic dynamics are discussed by Lind and Marcus [85].

[(3)](Page 13) This inequality, and the converse result that if a list of integers $\ell_1, \ell_2 \dots$ satisfies the inequality (1.5) then there is a prefix code with $\ell_i = |s(i)|$, was obtained by Kraft [78] and McMillan [95].

[(4)](Page 16) This seems to have first been proved by Fekete [41, p. 233] (in multiplicative form); a more accessible source is Pólya and Szegő [117, Chap. 3, Sect. 1].

[(5)](Page 23) A transformation which does not separate points widely or moves points around in a very orderly way has zero entropy, but it is important to understand that there is definitely no sense in which the converse holds. That is, there are transformations with zero entropy of great complexity.

[(6)](Page 30). The theory described in this section is due to Ornstein, and it is outlined in his monograph [103]. An elegant treatment using joinings may be found in the monograph of Rudolph [125].

[(7)](Page 30) In fact Mešalkin's result is more general, requiring only that the state probabilities each be of the form $\frac{a}{p^k}$ for some prime p and $a \in \mathbb{N}$ (and, by Theorem 1.19, the additional necessary condition that the two shifts have the same entropy).

[(8)](Page 32) This question was answered in the thesis of Polit [116], who constructed a pair of weakly isomorphic transformations of zero entropy that are not isomorphic. Rudolph [124] gave a more general approach to constructing examples of this kind, and for finding counterexamples to other natural conjectures. Other examples of weakly isomorphic systems were found by Thouvenot [137] using Gaussian processes, and by Lemańczyk [82] using product cocycles. More recently, Kwiatkowski, Lemańczyk, and Rudolph [79] have constructed weakly isomorphic C^∞ volume-preserving diffeomorphisms of $\mathbb{T}^2$ that are not isomorphic.

[(9)](Page 34) Let $\mathfrak{X}$ denote a subset of the set of all invertible measure-preserving transformations of a Borel probability space, with $\sim$ the equivalence relation of measurable isomorphism. A classifying space C is one for which there is a (reasonable) injective map $\mathfrak{X}/\sim \to C$; Ornstein's isomorphism theorem constructs such a map with $C = \mathbb{R}^+$ when $\mathfrak{X}$ is the class of Bernoulli shifts, while the Halmos–von Neumann theorem (see [38, Th. 6.13]) shows that C may be taken to be the set of all countable subgroups of $\mathbb{S}^1$ when $\mathfrak{X}$ is the class of transformations with discrete spectrum. Feldman [42] interpreted a construction of many mutually non-isomorphic K-automorphisms by Ornstein and Shields [106] to show that C certainly cannot be taken to be $\mathbb{R}^+$ when $\mathfrak{X}$ is the class of K automorphisms (a measure-preserving system $(X, \mathscr{B}, \mu, T)$ is called a K-automorphism if $h_\mu(T, \xi) > 0$ for any partition ξ with $H_\mu(\xi) > 0$; these systems have no zero-entropy factors). More recently, Foreman and Weiss [45] have used Hjorth's theory of turbulent equivalence relations [60] to show that C cannot be taken to be the collection of all isomorphism classes of countable groups when $\mathfrak{X}$ is the set of all invertible measure-preserving transformations.

2

Entropy for Continuous Maps

Measure-theoretic entropy was introduced to measure the complexity of measure-preserving transformations, as an invariant of measurable isomorphism, and to characterize distinguished measures. A continuous map T on a compact metric space has at least one invariant measure, so we start by analyzing the behavior of measure-theoretic entropy $h_\mu(T)$ as a function of the invariant probability measure $\mu \in \mathscr{M}^T$, where $\mathscr{M}^T$ denotes the space of T-invariant Borel probability measures on X. This problem is considered in Section 2.1, then in Section 2.2 a more intrinsic (purely topological) notion of entropy for continuous maps is introduced.

2.1 Continuity Properties of Entropy

Suppose that (X, d) is a compact metric space and $T : X \to X$ is a continuous map. Then the map

$$h : \mu \mapsto h_\mu(T)$$

is a map from $\mathscr{M}^T(X)$ to $[0, \infty]$. Recall from the discussion on [38, p. 454] that the space $\mathscr{M}^T(X)$ is weak*-compact. The next example shows that even for the simplest of maps T, the map h is not (lower semi-)continuous.

Example 2.1. Let $X = \mathbb{T}$ and $T(x) = 2x \pmod 1$. Then $m_\mathbb{T}$, Lebesgue measure on $\mathbb{T}$ is T-invariant. Define a sequence of T-invariant probability measures (μ_k) by

$$\mu_k = \frac{1}{3^k} \sum_{\ell=0}^{3^k-1} \delta_{\ell/3^k}.$$

Then $h_{\mu_k}(T) = 0$ for all k (since $(\mathbb{T}, \mathscr{B}, \mu_k, T)$ is measurably isomorphic to a permutation of 3^k points), while $h_{m_\mathbb{T}}(T) = \log 2$. Thus $\mu_k \to m_\mathbb{T}$ in the weak*-topology, but $h_{\mu_k}(T)$ does not converge to $h_{m_\mathbb{T}}(T)$. It follows that the entropy of a continuous map with respect to a weak*-convergent sequence of probability measures can jump up in the limit.

Example 2.2. In general $\mu \mapsto h_\mu(T)$ is also not upper semi-continuous. Let $X = \{(x, y) \mid 0 \leqslant y \leqslant x \leqslant 1\}$ and define a map T on X by

$$T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x & (\mathrm{mod}\ 1) \\ 2y & (\mathrm{mod}\ x) \end{pmatrix}.$$

For each $x \in [0, 1]$ let $\mu_x = \delta_x \times m_{[0,x]}$ where $m_{[0,x]}$ is Lebesgue measure on $[0, x]$ normalized to have $m_{[0,x]}([0, x]) = 1$. Then

$$h_{\mu_x}(T) = \begin{cases} 0 & \text{if } x = 0; \\ \log 2 & \text{if } x > 0, \end{cases}$$

while $\mu_x \to \mu_0$ as $x \to 0$ by definition. Thus the entropy of a continuous map with respect to a weak*-convergent sequence of probability measures can jump down in the limit.

Thus the entropy map $\mu \mapsto h_\mu(T)$ does not respect the topological structure of $\mathscr{M}^T(X)$. As we saw in Section 4.4, it does nonetheless respect the convex structure in the following sense. Whenever $\mu = \int_Z \mu_z \, \mathrm{d}\nu(z)$ is a disintegration of μ into T-invariant measures μ_z parameterized by a Borel probability space (Z, ν), we have

$$h_\mu(T) = \int h_{\mu_z}(T) \, \mathrm{d}\nu.$$

In particular, taking $Z = \{\mu_1, \mu_2\}$ with $\nu(\{\mu_1\}) = s$ and $\nu(\{\mu_2\}) = 1 - s$, we have the following theorem.

Theorem 2.3. *Let (X, d) be a compact metric space and let $T : X \to X$ be a continuous map. Then the map $h : \mathscr{M}^T(X) \to [0, \infty]$ is affine. This means that for any $s \in [0, 1]$ and measures $\mu_1, \mu_2 \in \mathscr{M}^T(X)$,*

$$h_{s\mu_1 + (1-s)\mu_2}(T) = s h_{\mu_1}(T) + (1 - s) h_{\mu_2}(T).$$

Notice that Example 2.1 is a quite natural dynamical system and sequence of weakly converging measures, while Example 2.2 seems artificially constructed to fail to be upper semi-continuous. The next definition introduces a natural geometric condition[(10)] that excludes this behavior.

Definition 2.4. *A homeomorphism $T : (X, \mathsf{d}) \to (X, \mathsf{d})$ of a compact metric space is called* expansive *if there is a $\delta > 0$ such that*

$$\sup_{n \in \mathbb{Z}} \mathsf{d}(T^n x, T^n y) \leqslant \delta \implies x = y. \tag{2.1}$$

Any $\delta > 0$ that has the property in equation (2.1) is called an *expansive constant* for T.

Example 2.5. The automorphism of $\mathbb{T}^2$ associated to the matrix $\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$ from Section 4.5 and [38, Ex. 1.34] is expansive.

More generally, an automorphism of $\mathbb{T}^k$ associated to a matrix A is expansive if and only if the matrix A is hyperbolic, which means the eigenvalues all have modulus not equal to one[(11)].

Lemma 2.6. *Let $T : (X, \mathsf{d}) \to (X, \mathsf{d})$ be an expansive homeomorphism with expansive constant δ. Then any finite partition ξ with the property that $\operatorname{diam}(P) < \delta$ for all $P \in \xi$ is a generator of the measure-preserving system $(X, \mathscr{B}, \mu, T)$ for any $\mu \in \mathscr{M}^T$.*

Proof. Let

$$\mathscr{A}_0 = \bigcup_{n=1}^{\infty} \bigvee_{k=-n}^{n} T^{-k}\xi,$$

which is a countable algebra. We need to show that $\mathscr{A}_0$ generates the σ-algebra $\mathscr{B}$. For this it is enough to show that any open set $O \subseteq X$ can be written as a union (automatically a countable union) of elements of $\mathscr{A}_0$.

Let $O \subseteq X$ be open, and let $x \in O$. Then we claim that there exists some n with

$$[x]_{\bigvee_{k=-n}^{n} T^{-k}\xi} \subseteq O.$$

Suppose the opposite. Then for every n there would exist some $y_n \notin O$ with $\mathsf{d}(T^k x, T^k y) < \delta$ for $k = -n, \ldots, n$. Let y be the limit of a convergent subsequence of the sequence (y_n), then continuity of T shows that $\mathsf{d}(T^k x, T^k y) < \delta$ for any $k \in \mathbb{Z}$. The claim implies the lemma. □

In contrast to Example 2.2, for an expansive map the entropy can jump up but cannot jump down at a weak*-limit.

Theorem 2.7. *If $T : (X, \mathsf{d}) \to (X, \mathsf{d})$ is an expansive homeomorphism, then the map $\mu \mapsto h_\mu(T)$ is upper semi-continuous.*

What this means is that for any measure μ and for any $\varepsilon > 0$, there is a weak*-neighborhood U of μ with the property that

$$\nu \in U \implies h_\nu(T) < h_\mu(T) + \varepsilon.$$

For any set $B \subseteq X$, a metric space, write B^o for the interior of B, $\overline{B}$ for the closure and ∂B for the boundary.

Lemma 2.8. *If $\mu(\partial B) = 0$ then $\mu_n \to \mu$ weak* implies that $\mu_n(B) \to \mu(B)$.*

Proof. Since $\mu(\partial B) = 0$, for any $\varepsilon > 0$ there is a compact set K and an open set U with $\mu(U \setminus K) < \varepsilon$, $B^o \underset{\mu}{=} \overline{B}$, and $K \subseteq B^o \subseteq B \subseteq \overline{B} \subseteq U$. Choose Urysohn functions f and g with $\chi_K < f < \chi_{B^o}$ and $\chi_{\overline{B}} < g < \chi_U$.

Then $\int f\,\mathrm{d}\mu \leqslant \mu(B) \leqslant \int g\,\mathrm{d}\mu$ and $\int (g-f)\,\mathrm{d}\mu < \varepsilon$. Now by the weak*-convergence, there is an N such that if $k > N$ then

$$\left| \int f\,\mathrm{d}\mu - \int f\,\mathrm{d}\mu_k \right| < \varepsilon$$

and

$$\left| \int g\,\mathrm{d}\mu - \int g\,\mathrm{d}\mu_k \right| < \varepsilon.$$

It follows that for $k > N$,

$$\int f\,\mathrm{d}\mu - \varepsilon < \int f\,\mathrm{d}\mu_k < \mu_k(B) < \int g\,\mathrm{d}\mu_k < \int g\,\mathrm{d}\mu + \varepsilon,$$

so $|\mu(B) - \mu_k(B)| < 2\varepsilon$. □

Proof of Theorem 2.7. Choose a partition $\xi = \{P_1, \ldots, P_k\}$ with

$$\operatorname{diam}(P_i) < \delta, \mu\left(\partial P_i\right) = 0$$

for all $i \leqslant k$, where δ is an expansive constant for T. To construct such a partition, choose for each $x \in X$ an $\varepsilon_x \in (0, \delta/2)$ with $\mu\left(\partial B_{\varepsilon_x}(x)\right) = 0$; the open cover

$$\{B_{\varepsilon_x}(x) \mid x \in X\}$$

has a finite subcover from which ξ may be constructed. By Lemma 2.6 the partition ξ is a generator. Fix $\varepsilon > 0$ and choose N with

$$\frac{1}{N} H_\mu \left(\bigvee_{j=0}^{N-1} T^{-j}\xi \right) < h_\mu(T) + \varepsilon.$$

The partition $\bigvee_{j=0}^{N-1} T^{-j}\xi$ contains finitely many atoms Q each with the property that $\mu\left(\partial Q\right) = 0$. By Lemma 2.8 there is a weak*-neighborhood U of μ with the property that

$$\nu \in U \implies |\nu(Q) - \mu(Q)| < \varepsilon' \text{ for all } Q \in \bigvee_{j=0}^{N-1} T^{-j}\xi.$$

It follows, by continuity of ϕ, that if ε' is small enough,

$$\frac{1}{N} H_\nu \left(\bigvee_{j=0}^{N-1} T^{-j}\xi \right) \leqslant \frac{1}{N} H_\mu \left(\bigvee_{j=0}^{N-1} T^{-j}\xi \right) + \varepsilon \leqslant h_\mu(T) + 2\varepsilon,$$

so $h_\nu(T) \leqslant h_\mu(T) + 2\varepsilon$, since

$$h_\nu(T) = \inf_{n \in \mathbb{N}} \left\{ \frac{1}{n} H_\nu \left(\bigvee_{j=0}^{n-1} T^{-j}\xi \right) \right\}.$$

□

It is not difficult to generalize Definition 2.4 through Theorem 2.7 to give the notion of *forwardly expansive maps* and the result that the entropy map $\mu \mapsto h_\mu(T)$ is upper semi-continuous when T is a forwardly expansive map. Using this, Example 2.1 is a forwardly expansive endomorphism of $\mathbb{T}$. This shows that expansiveness cannot give more than upper semi-continuity (see also Exercises 2.1.1 and 2.1.2).

Exercises for Section 2.1

Exercise 2.1.1. Show that $\mu \mapsto h_\mu(\sigma)$ is not lower semi-continuous for the full shift $\sigma : \{0, \ldots, k-1\}^{\mathbb{Z}} \to \{0, \ldots, k-1\}^{\mathbb{Z}}$ on an alphabet with $k \geqslant 2$ symbols.

Exercise 2.1.2. Show that $\mu \mapsto h_\mu(T_A)$ is not lower semi-continuous for the automorphism T_A of $\mathbb{T}^2$ associated to the matrix $A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$ from Section 4.5.

Exercise 2.1.3. Let $T : X \to X$ be an expansive homeomorphism of a compact metric space. Show that for any $n \geqslant 1$ the set $\{x \in X \mid T^n x = x\}$ of points of period n under T is finite.

2.2 Topological Entropy

Entropy was originally defined purely in terms of measure-theoretic properties. A continuous map may have many invariant measures, so it became important to define a topological analog of entropy to reflect topological properties of a map. There are two ways in which the dependence on a specific measure may be removed in defining the entropy of a continuous map. First, the set

$$\{h_\mu(T) \mid \mu \in \mathscr{M}^T\}$$

may be considered as a whole. Since entropy is an affine function on the convex set $\mathscr{M}^T$ (see Theorem 2.3), this set is an interval in $[0, \infty]$, so it is natural to associate the quantity $\sup\{h_\mu(T) \mid \mu \in \mathscr{M}^T\}$ to T. Second, one may try to emulate the definition of measure-theoretic entropy using the metric or topological structure of a continuous map.

The second approach, which we develop now, was initiated by Adler, Konheim and McAndrew [3]. We shall see later that the two notions end up with the same quantity.

Definition 2.9. *For a cover $\mathscr{U}$ of a compact topological space X, define $N(\mathscr{U})$ to be the smallest cardinality of a subcover of $\mathscr{U}$, and define the entropy of $\mathscr{U}$ to be $H(\mathscr{U}) = \log N(\mathscr{U})$.*

For covers $\mathscr{U} = \{U_i\}_{i\in I}$ and $\mathscr{V} = \{V_j\}_{j\in J}$, write $\mathscr{U} \vee \mathscr{V}$ for the cover

$$\mathscr{U} \vee \mathscr{V} = \{U_i \cap V_j\}_{i\in I, j\in J}.$$

Notice that if T is continuous, then for any open cover $\mathscr{U}$, $T^{-1}(\mathscr{U})$ is also an open cover. Moreover, the sequence (a_n) defined by

$$a_n = \log N\left(\bigvee_{i=0}^{n-1} T^{-i}\mathscr{U}\right)$$

is subadditive. This proves the convergence statement implicit in Definition 2.10, by Lemma 1.13.

Definition 2.10. *Let $T : X \to X$ be a continuous map on a compact topological space. The* cover entropy *of T with respect to an open cover $\mathscr{U}$ is defined to be*

$$h_{\text{cover}}(T, \mathscr{U}) = \lim_{n\to\infty} \frac{1}{n} \log N\left(\bigvee_{i=0}^{n-1} T^{-i}\mathscr{U}\right) = \inf_{n\geqslant 1} \frac{1}{n} \log N\left(\bigvee_{i=0}^{n-1} T^{-i}\mathscr{U}\right),$$

and the cover entropy *of T is defined to be*

$$h_{\text{cover}}(T) = \sup_{\mathscr{U}} h_{\text{cover}}(T, \mathscr{U})$$

where the supremum is taken over all open covers of X.

As was the case with measure-theoretic entropy, it is not clear how to use this definition to actually compute topological entropy. The first step towards solving this problem comes from the existence of a *generator*. The results below make sense for continuous maps on compact spaces, but we restrict attention to the metric setting since that is simpler and covers what is needed for the applications considered here.

Definition 2.11. *Let $T : (X, \mathsf{d}) \to (X, \mathsf{d})$ be a continuous map on a compact metric space. A one-sided generator for T is a finite open cover $\mathscr{U} = \{U_i\}_{i\in I}$ with the property that for any sequence $(i_k)_{k\geqslant 0}$, the set*

$$\bigcap_{k\geqslant 0} T^{-k}\left(\overline{U_{i_k}}\right)$$

contains at most a single point.

In order to work with open covers, we need a standard result about compact metric spaces.

Lemma 2.12. *Let (X, d) be a compact metric space and $\mathscr{U}$ an open cover of X. Then there is some $\delta > 0$, called a* Lebesgue number *for $\mathscr{U}$, with the property that any subset of X with diameter less than δ is contained in some element of $\mathscr{U}$.*

PROOF. If no Lebesgue number exists, then there is an open cover $\mathscr{U}$ such that for any $\delta > 0$ there is an $x \in X$ with the property that no $U \in \mathscr{U}$ contains $B_\delta(x)$. Thus for each $n \geqslant 1$ we can choose a sequence $(x_n) \subseteq X$ such that $B_{1/n}(x_n) \nsubseteq$ for any $U \in \mathscr{U}$. By compactness, we may choose a convergent subsequence (x_{n_j}) with $x_{n_j} \to x^* \in X$ as $j \to \infty$. Since $\mathscr{U}$ is an open cover, there is some $\delta > 0$ and some $U^* \in \mathscr{U}$ with $B_\delta(x^*) \subseteq U^*$. Choose n so that $\frac{1}{n} < \frac{\delta}{2}$ and $\mathsf{d}(x_{n_j}, x^*) < \frac{\delta}{2}$ for $n_j > n$. Then $B_{1/n}(x_{n_j}) \subseteq U^*$, which is a contradiction. □

Theorem 2.13. *Let $\mathscr{U}$ be a one-sided generator for a continuous map*

$$T : (X, \mathsf{d}) \to (X, \mathsf{d})$$

on a compact metric space. Then $h_{\mathrm{cover}}(T, \mathscr{U}) = h_{\mathrm{cover}}(T)$.

PROOF. We claim first that for any $\varepsilon > 0$ there is an M such that each set in the refinement

$$\bigvee_{j=0}^{M} T^{-j}\mathscr{U}$$

has diameter less than or equal to ε. If not, there is some $\varepsilon > 0$ such that for each $k \geqslant 0$ we may find points $x_1^{(k)}, x_2^{(k)}$ with $\mathsf{d}(x_1^{(k)}, x_2^{(k)}) > \varepsilon$ that lie in the same element of $\bigvee_{n=0}^{k} T^{-n}\mathscr{U}$. By compactness, there is a subsequence k_j with $x_1^{(k_j)} \to x_1^*$ and $x_2^{(k_j)} \to x_2^*$, and $\mathsf{d}(x_1^*, x_2^*) \geqslant \varepsilon$. By construction, the points x_1^* and x_2^* lie in a single set

$$\bigcap_{k \geqslant 0} T^{-k}\left(\overline{U_{i_k}}\right)$$

for some sequence (i_k), contradicting the assumption that $\mathscr{U}$ is a one-sided generator.

Now let $\mathscr{V}$ be any open cover of X, and let δ be a Lebesgue number for $\mathscr{V}$. By the claim, there is some M for which each set in $\bigvee_{n=0}^{M} T^{-n}\mathscr{U}$ has diameter less than δ, so $\mathscr{V}$ is refined by $\bigvee_{n=0}^{M} T^{-n}\mathscr{U}$ in the sense that for every $x \in X$ and every $V \in \mathscr{V}$ with $x \in V$ there exists an element $U_M \in \bigvee_{j=0}^{M} T^{-j}\mathscr{U}$ with

$$x \in U_M \subseteq V.$$

The same holds for the refinement $\bigvee_{j=0}^{k-1} T^{-j}\mathscr{V}$ and $\bigvee_{n=0}^{M} T^{-n}\mathscr{U}$. This implies that

$$\begin{aligned} h_{\mathrm{cover}}(T,\mathscr{V}) &\leqslant h_{\mathrm{cover}}\left(T,\bigvee_{n=0}^{M}T^{-n}\mathscr{U}\right)\\ &= \lim_{k\to\infty}\frac{1}{k}\log N\left(\bigvee_{j=0}^{k-1}T^{-j}\bigvee_{n=0}^{M}T^{-n}\mathscr{U}\right)\\ &= \lim_{k\to\infty}\frac{1}{k}\log N\left(\bigvee_{j=0}^{M+k-1}T^{-j}\mathscr{U}\right)\\ &= \lim_{k\to\infty}\frac{M+k-1}{k}\frac{1}{M+k-1}\log N\left(\bigvee_{j=0}^{M+k-1}T^{-j}\mathscr{U}\right)\\ &= h_{\mathrm{cover}}(T,\mathscr{U}), \end{aligned}$$

so $h_{\mathrm{cover}}(T) = h_{\mathrm{cover}}(T,\mathscr{U})$ since this holds for any cover $\mathscr{V}$. □

Corollary 2.14. *Let X be a closed, shift-invariant subset of the one-sided full shift $\prod_{n\geqslant 0}\{0,1,\ldots,s-1\}$, and let $t_n(X)$ denote the number of blocks of length n that appear in any element of X. Then $\frac{1}{n}\log t_n(X)$ converges, and*

$$h_{\mathrm{cover}}(\sigma|_X) = \lim_{n\to\infty}\frac{1}{n}\log t_n(X).$$

Proof. The open cover $\mathscr{U}$ by the sets $\{x\in X \mid x_0 = j\}$ for $j\in\{0,1,\ldots,k-1\}$ is a generator, and $t_n(X) = N\left(\bigvee_{j=0}^{n-1}T^{-j}\mathscr{U}\right)$; the result follows by Theorem 2.13. □

Theorem 2.13 is a direct analog of Theorem 1.20; it is also useful to have the analog of Exercise 1.3.1.

Proposition 2.15. *Let $T : X\to X$ be a continuous map on a compact metric space (X,d). If $(\mathscr{U}_n)$ is a sequence of open covers of X with $\operatorname{diam}(\mathscr{U}_n)\to 0$ as $n\to\infty$, then*

$$\lim_{n\to\infty} h_{\mathrm{cover}}(T,\mathscr{U}_n) = h_{\mathrm{cover}}(T)$$

(with the convention that if $h_{\mathrm{cover}}(T)=\infty$ then $h_{\mathrm{cover}}(T,\mathscr{U}_n)$ diverges to ∞).

Proof. Assume first that $h_{\mathrm{cover}}(T)$ is finite, fix $\varepsilon>0$, choose an open cover $\mathscr{V}$ for which $h_{\mathrm{cover}}(T,\mathscr{V}) > h_{\mathrm{cover}}(T)-\varepsilon$, and let δ be a Lebesgue number for $\mathscr{V}$. Choose N so that $n\geqslant N$ implies that $\operatorname{diam}(\mathscr{U}_n)<\delta$, so $\mathscr{U}_n$ refines $\mathscr{V}$ and hence

$$h_{\mathrm{cover}}(T)\geqslant h_{\mathrm{cover}}(T,\mathscr{U}_n) > h_{\mathrm{cover}}(T)-\varepsilon,$$

showing that $\lim_{n\to\infty} h_{\mathrm{cover}}(T,\mathscr{U}_n) = h_{\mathrm{cover}}(T)$. If $h_{\mathrm{cover}}(T)=\infty$ then for any R there is an open cover $\mathscr{V}$ for which $h_{\mathrm{cover}}(T,\mathscr{V})>R$, and we may proceed as before. □

In topological dynamics, the behavior of individual points is visible, and this allows an alternative approach to entropy via measuring the asymptotic complexity of the metric space with respect to a sequence of metrics defined using the continuous map (equivalently, measuring the complexity of the orbits under the map). This approach is due to Dinaburg [31] and Bowen; as we shall see in Section 3.3 Bowen in particular used this to extend the definition of topological entropy to uniformly continuous maps on locally compact metric spaces.

Definition 2.16. *Let $T : (X, \mathsf{d}) \to (X, \mathsf{d})$ be a continuous map on a compact metric space. A subset $F \subseteq X$ is* (n, ε)-spanning *if for every $x \in X$ there is a point $y \in F$ with*

$$\mathsf{d}(T^i x, T^i y) \leqslant \varepsilon \text{ for } i = 0, \ldots, n-1.$$

A subset $E \subseteq X$ is (n, ε)-separated *if for any two distinct points $x, y \in E$,*

$$\max_{0 \leqslant i \leqslant n} \mathsf{d}(T^i x, T^i y) > \varepsilon.$$

Define $r_n(\varepsilon)$ to be the smallest cardinality of an (n, ε)-spanning set, and define $s_n(\varepsilon)$ to be the largest cardinality of an (n, ε)-separated set. Notice that both are finite by compactness. The spanning set entropy *of T is*

$$h_{\mathrm{span}}(T) = \lim_{\varepsilon \searrow 0} \limsup_{n \to \infty} \frac{1}{n} \log r_n(\varepsilon).$$

The separated set entropy *of T is*

$$h_{\mathrm{sep}}(T) = \lim_{\varepsilon \searrow 0} \limsup_{n \to \infty} \frac{1}{n} \log s_n(\varepsilon).$$

Notice that

$$\varepsilon < \varepsilon' \implies r_n(\varepsilon) \geqslant r_n(\varepsilon') \text{ and } s_n(\varepsilon) \geqslant s_n(\varepsilon'),$$

so the limit in ε exists in both cases.

Lemma 2.17.

$$r_n(\varepsilon) \leqslant s_n(\varepsilon) \leqslant r_n(\varepsilon/2).$$

Proof. If E is an (n, ε)-separated set of maximal cardinality (so $|E| = s_n(\varepsilon)$), then it is also (n, ε)-spanning since if it is not (n, ε)-spanning another point could be added to make an (n, ε)-separated set of larger cardinality. Thus $r_n(\varepsilon) \leqslant |E| = s_n(\varepsilon)$.

If F is an $(n, \varepsilon/2)$-spanning set of minimal cardinality (so $|F| = r_n(\varepsilon/2)$) and $x \in F$, then there can only be one point $y \in E$ (where E is again a maximal (n, ε)-separated set) with

$$\mathsf{d}(T^i x, T^i y) \leqslant \frac{\varepsilon}{2} \text{ for } i = 0, \ldots, n.$$

This is because if $y_1, y_2 \in E$ both satisfy this estimate, then $\mathsf{d}(T^i y_i, T^i y_2) \leqslant \varepsilon$ for $i = 0, \ldots, n-1$, which implies that $y_1 = y_2$ since E is (n, ε)-separated. It follows that $|E| = s_n(\varepsilon) \leqslant |F| = r_n(\varepsilon/2)$. □

Theorem 2.18. *For a continuous map $T : (X, \mathsf{d}) \to (X, \mathsf{d})$ on a compact metric space,*

$$h_{\mathrm{cover}}(T) = h_{\mathrm{sep}}(T) = h_{\mathrm{span}}(T).$$

The common value is called the topological entropy *of T, denoted $h_{\mathrm{top}}(T)$.*

Proof. Lemma 2.17 shows that $h_{\mathrm{sep}}(T) = h_{\mathrm{span}}(T)$.

Let $\mathscr{U}$ be an open cover with Lebesgue number ε, and let F be an $(n, \varepsilon/2)$-spanning set of cardinality $r_n(\varepsilon/2)$. For $R \geqslant 0$ and $x \in X$ we write

$$\overline{B}_R(x) = \{y \in X \mid \mathsf{d}(x, y) \leqslant R\}$$

for the closed ball of radius R around x. Then

$$X = \bigcup_{x \in F} \bigcap_{i=0}^{n-1} T^{-i} \overline{B}_{\varepsilon/2}(T^i x).$$

Since the Lebesgue number of $\mathscr{U}$ is ε, for every x and i there is some $U \in \mathscr{U}$ with

$$\overline{B}_{\varepsilon/2}(T^i x) \subseteq B_\varepsilon(T^i x) \subseteq U.$$

Choosing the elements $\bigcap T^{-i} U_i$ of the cover $\bigvee_{i=0}^{n-1} T^{-i}\mathscr{U}$ for all $x \in F$ shows that

$$N\left(\bigvee_{i=0}^{n-1} T^{-i}\mathscr{U}\right) \leqslant r_n(\varepsilon/2). \tag{2.2}$$

Now let $\mathscr{U}$ be an open cover with $\operatorname{diam}(U) \leqslant \varepsilon$ for all $U \in \mathscr{U}$, and let E be an (n, ε)-separated set with cardinality $s_n(\varepsilon)$. Then no member of

$$\mathscr{V} = \bigvee_{i=0}^{n-1} T^{-i}\mathscr{U}$$

can contain two elements of E, so any subcover of $\mathscr{V}$ must have at least $s_n(\varepsilon)$ elements. Thus

$$s_n(\varepsilon) \leqslant N\left(\bigvee_{i=0}^{n-1} T^{-i} U\right). \tag{2.3}$$

Equation (2.2) implies that $h_{\mathrm{span}}(T) \geqslant h_{\mathrm{cover}}(T)$, while equation (2.3) implies that $h_{\mathrm{sep}}(T) \leqslant h_{\mathrm{cover}}(T)$. □

This result means that we can speak about 'topological entropy' and use whichever definition is most convenient for continuous maps on compact metric spaces.

Lemma 2.19. *If $T : X \to X$ is a homeomorphism of a compact metric space, then $h_{\text{top}}(T^{-1}) = h_{\text{top}}(T)$.*

PROOF. Notice that for any continuous map T and open cover $\mathscr{U}$ we have

$$N(T^{-1}\mathscr{U}) \leqslant N(\mathscr{U}),$$

but if T is a homeomorphism we have $N(T^{-1}\mathscr{U}) = N(\mathscr{U})$. Thus

$$\begin{aligned} h_{\text{cover}}(T, \mathscr{U}) &= \lim_{n\to\infty} \frac{1}{n} \log N \left(\bigvee_{i=0}^{n-1} T^{-i}\mathscr{U} \right) \\ &= \lim_{n\to\infty} \frac{1}{n} \log N \left(T^{n-1} \bigvee_{i=0}^{n-1} T^{-i}\mathscr{U} \right) \\ &= \lim_{n\to\infty} \frac{1}{n} \log N \left(\bigvee_{i=0}^{n-1} T^{i}\mathscr{U} \right) \\ &= h_{\text{cover}}(T^{-1}, \mathscr{U}). \end{aligned}$$

□

Lemma 2.20. *If $T_i : X_i \to X_i$ are continuous maps of compact metric spaces for $i = 1, 2$, then $h_{\text{top}}(T_1 \times T_2) = h_{\text{top}}(T_1) + h_{\text{top}}(T_2)$.*

PROOF. For open covers of $X_1 \times X_2$ of the product form

$$\mathscr{U} \times \mathscr{V} = \{U \times V \mid U \in \mathscr{U}, V \in \mathscr{V}\}$$

we have

$$N^{X_1\times X_2}(\mathscr{U} \times \mathscr{V}) = N^{X_1}(\mathscr{U}) \times N^{X_2}(\mathscr{V})$$

and

$$(\mathscr{U} \times \mathscr{V}) \vee (\mathscr{U}' \times \mathscr{V}') = (\mathscr{U} \vee \mathscr{U}') \times (\mathscr{V} \vee \mathscr{V}'),$$

so

$$h_{\text{cover}}(T_1 \times T_2, \mathscr{U} \times \mathscr{V}) = h_{\text{cover}}(T_1, \mathscr{U}) + h_{\text{cover}}(T_2, \mathscr{V}),$$

and therefore $h_{\text{top}}(T_1 \times T_2) \geqslant h_{\text{top}}(T_1) + h_{\text{top}}(T_2)$.

For the reverse inequality we need to check that these product covers do enough. Every open set in $X_1 \times X_2$ is a union of open rectangles $U \times V$, so for any open cover $\mathscr{O}$ of $X_1 \times X_2$ we can choose a refinement that consists only of open rectangles; choose from this a minimal subcover

$$\mathscr{O}' = \{U'_1 \times V'_1, \ldots, U'_{N(\mathscr{O}')} \times V'_{N(\mathscr{O}')}\} \geqslant \mathscr{O}$$

(here $\geqslant$ means refinement). Let

$$\mathscr{U}' = \{U'_1, \ldots, U'_{N(\mathscr{O}')}\}$$

and

$$\mathscr{V}' = \{V'_1, \ldots, V'_{N(\mathscr{O}')}\},$$

and for a given point $(x^1, x^2) \in X_1 \times X_2$ let

$$U^{(x^1)} = \bigcap_{U'_i \ni x} U'_i$$

and

$$V^{(x^2)} = \bigcap_{V'_i \ni x} V'_i.$$

These sets are open, so by compactness we may choose points $x^1_1, \ldots, x^1_m \in X_1$ and $x^2_1, \ldots, x^2_n \in X_2$ for which

$$\mathscr{U} = \{U^{(x^1_1)}, \ldots, U^{(x^1_m)}\}$$

and

$$\mathscr{V} = \{V^{(x^2_1)}, \ldots, V^{(x^2_n)}\}$$

are open covers of X_1 and X_2 respectively. Now consider any set

$$U^{(x^1_i)} \times V^{(x^2_j)} \in \mathscr{U} \times \mathscr{V}.$$

Since $\mathscr{O}'$ covers $X_1 \times X_2$, we have $(x^1_i, x^2_j) \in U'_k \times V'_k$ for some k, $1 \leqslant k \leqslant N(\mathscr{O}')$. By construction, $U^{(x^1_i)} \subseteq U'_k$ and $V^{(x^2_j)} \subseteq V'_k$, so $U^{(x^1_i)} \times V^{(x^2_j)} \subseteq U'_k \times V'_k$, which implies that $\mathscr{O} \leqslant \mathscr{O}' \leqslant \mathscr{U} \times \mathscr{V}$, and therefore

$$h_{\mathrm{cover}}(T_1 \times T_2, \mathscr{O}) \leqslant h_{\mathrm{cover}}(T_1, \mathscr{U}) + h_{\mathrm{cover}}(T_2, \mathscr{V}),$$

completing the proof. □

Exercises for Section 2.2

Exercise 2.2.1. Show that Corollary 2.14 also holds for two-sided shifts.

Exercise 2.2.2. Let $T_i : X_i \to X_i$ be a continuous map of a compact metric space for $i = 1, 2$. Prove that if (X_2, T_2) is a topological factor of (X_1, T_1) (that is, there is a continuous map $\theta : X_1 \to X_2$ with $\theta \circ T_1 = T_2 \circ \theta$), then

$$h_{\mathrm{top}}(T_2) \leqslant h_{\mathrm{top}}(T_1).$$

Exercise 2.2.3. Let $T : X \to X$ be a continuous map on a compact metric space. Prove that $h_{\mathrm{top}}(T^k) = k h_{\mathrm{top}}(T)$ for any $k \geqslant 1$ using the definition in terms of spanning and separating sets (this will also be proved later as Corollary 3.6).

Exercise 2.2.4. Show that an expansive homeomorphism of a compact metric space has finite topological entropy.

2.3 The Variational Principle

The variational principle[12] expresses a relationship between the topological entropy of a continuous map and the entropy with respect to invariant measures.

2.3.1 Periodic Points Producing Positive Entropy

To motivate the proof of Theorem 2.22 in Section 2.3.2 we state and prove a simpler result regarding the limiting behavior of periodic orbits for the map $T_p : x \mapsto px \pmod 1$ on the circle $\mathbb{T}$. As we will see later in the proof of Theorem 2.22, the argument relies on the discrete distribution and spacing properties of a sequence of measures before a weak*-limit is taken. The result is a simple version of a similar phenomena used by Einsieder, Lindenstrauss, Michel and Venkatesh [36].

Proposition 2.21. *Let $T_p(x) = px \mod 1$ for $x \in \mathbb{T}$ for some $p \geqslant 2$. Let $\alpha > 0$ and $C > 0$, and assume that for every n with $\gcd(n,p) = 1$ we choose a subset $S_n \subseteq \{0, \frac{1}{n}, \ldots, \frac{n-1}{n}\}$ with $T_p(S_n) \subseteq S_n$ and $|S_n| \geqslant Cn^\alpha$. Let*

$$\mu_n = \frac{1}{|S_n|} \sum_{x \in S_n} \delta_x$$

be the normalized counting measure on S_n for $n \geqslant 1$. Then any weak-limit μ of (μ_n) is T_p-invariant and satisfies $h_\mu \geqslant \alpha \log p$.*

Proof. Let $\xi = \{[0, \frac{1}{p}), [\frac{1}{p}, \frac{2}{p}), \ldots, [\frac{p-1}{p}, 1)\}$ be the generator for $T = T_p$ from Example 1.27. Notice that

$$\mu(\{0\}) = \mu\big(T^{-1}\{0\}\big) = \mu\big(\{0\} \sqcup \{\tfrac{1}{p}, \tfrac{2}{p}, \ldots, \tfrac{p-1}{p}\}\big)$$

for any T-invariant measure μ. It follows that $\{\frac{1}{p}, \frac{2}{p}, \ldots, \frac{p-1}{p}\}$ is a μ-null set, and so $\mu(\partial P) = 0$ for all $P \in \xi$ unless $0 \in \partial P$ and $\mu(\{0\}) > 0$.

Assume first that $\mu_{n_k} \to \mu$ in the weak*-topology, and $\mu(\{0\}) = 0$. Then, for any fixed $m \geqslant 1$, we have $\mu(\partial P) = 0$ for all $P \in \bigvee_{i=0}^{m-1} T^{-i}\xi$ and so

$$\mu(P) = \lim_{k \to \infty} \mu_{n_k}(P)$$

by Lemma 2.8. Just as in the proof of Theorem 2.7, this implies that

$$H_\mu(\xi_m) = \lim_{k \to \infty} H_{\mu_{n_k}}(\xi_m) \tag{2.4}$$

for any $m \geqslant 1$, where we write $\xi_N = \bigvee_{i=0}^{N-1} T^{-i}\xi$ for $N \geqslant 1$.

For each $k \geqslant 1$ write $\lceil \log_p n_k \rceil = d_k m + r_k$ with $d_k \geqslant 0$ and $0 \leqslant r \leqslant m-1$. By subadditivity of entropy (Proposition 1.7) and T-invariance of μ_{n_k} we have

$$\begin{aligned} H_{\mu_{n_k}}\left(\xi_{\lceil \log_p n_k \rceil}\right) &\leqslant d_k H_{\mu_{n_k}}(\xi_m) + H_{\mu_{n_k}}(\xi_{r_k}) \\ &\leqslant d_k H_{\mu_{n_k}}(\xi_m) + m \log p. \end{aligned} \tag{2.5}$$

Notice that the atoms of the partition $\xi_{\lceil \log_p n_k \rceil}$ are intervals of length

$$\frac{1}{p} p^{-(\lceil \log_p n_k \rceil - 1)} \leqslant \frac{1}{n_k},$$

so that each such interval contains at most one member of S_{n_k}. In other words, as far as the measure μ_{n_k} is concerned, this partition is the partition into $|S_{n_k}|$ sets of equal measure (and many null sets), so that

$$H_{\mu_{n_k}}\left(\xi_{\lceil \log_p n_k \rceil}\right) = \log |S_{n_k}|.$$

Together with equation (2.5) we see that

$$\begin{aligned} \frac{1}{m} H_{\mu_{n_k}}(\xi_m) &\geqslant \frac{1}{m d_k} \log |S_{n_k}| - \frac{1}{d_k} \log p \\ &\geqslant \frac{\log_p n}{m d_k} \alpha \log p + \frac{1}{m d_k} \log C - \frac{1}{d_k} \log p, \end{aligned} \tag{2.6}$$

since, by assumption, $|S_{n_k}| \geqslant C n_k^{\alpha}$. As $k \to \infty$ we have $\frac{\log_p n_k}{m d_k} \to 1$ and $\frac{1}{d_k} \to 0$ by definition of d_k. By equation (2.4) we conclude that

$$\frac{1}{m} H_{\mu}(\xi_m) \geqslant \alpha \log p$$

for any $m \geqslant 1$, which shows that $h_{\mu}(T) \geqslant \alpha \log p$ and proves the result under the assumption that $\mu(\{0\}) = 0$.

Assume now that $\mu(\{0\}) > 0$, so that we cannot use equation (2.4) for the partition ξ_m. However, only two partition elements $P \in \xi_m$ fail to have the property $\mu(\partial P) = 0$, producing a bounded error only. Let $P_0 = [0, \frac{1}{p^m})$ and $P_{-1} = [\frac{p^m - 1}{p^m}, 1)$ be the two problematical elements of ξ_m, and define a new partition

$$\eta_m = \{P_0 \cup P_{-1}, \xi_m \setminus \{P_0, P_{-1}\}\}.$$

Notice that ξ_m is a refinement of η_m, since

$$\xi_m = \eta_m \vee \{P_0, P_{-1}\}.$$

Thus

$$\begin{aligned} H_{\mu_{n_k}}(\xi_m) &= H_{\mu_{n_k}}(\eta_m) + H_{\mu_{n_k}}\left(\{P_0, \mathbb{T} \setminus P_0\} \middle| \eta_m\right) \\ &\leqslant H_{\mu_{n_k}}(\eta_m) + \log 2. \end{aligned}$$

Together with equation (2.6) this gives

$$\frac{1}{m} H_{\mu_{n_k}}(\eta_m) \geqslant \frac{\log_p n_k}{m d_k} \alpha \log p - \frac{1}{d_k} \log p + \frac{1}{m d_k} \log C - \frac{1}{m} \log 2,$$

and we may use the weak*-convergence of μ_{n_k} since $\mu(\partial P) = 0$ for all $P \in \eta_m$. Therefore

$$\frac{1}{m} H_\mu(\xi_m) \geqslant \frac{1}{m} H_\mu(\eta_m) \geqslant \alpha \log p - \frac{1}{m} \log 2,$$

which again implies that $h_\mu(T) \geqslant \alpha \log p$. □

2.3.2 The Variational Principle

The first inequality in the proof of Theorem 2.22 was shown by Goodwyn [55]; Dinaburg [31] proved the full result under the assumption that X has finite covering dimension. Finally Goodman [53] proved the result without any assumptions. The first proof below is a simplification due to Misiurewicz [100], and the second is a strengthened version due to Blanchard, Glasner and Host [10]. Write $\mathscr{E}^T$ for the subset of $\mathscr{M}^T$ consisting of those Borel probability measures on X that are ergodic with respect to T.

Theorem 2.22 (Variational Principle). *Let $T : (X, \mathsf{d}) \to (X, \mathsf{d})$ be a continuous map on a compact metric space. Then*

$$\begin{aligned} h_{\mathrm{top}}(T) &= \sup_{\mu \in \mathscr{M}^T(X)} h_\mu(T) \\ &= \sup_{\mu \in \mathscr{E}^T(X)} h_\mu(T). \end{aligned}$$

A measure $\mu \in \mathscr{M}^T(X)$ with $h_{\mathrm{top}}(X) = h_\mu(T)$ is called a *maximal measure*. A maximal measure does not always exist (see Example 2.31), but does if T is expansive (see Corollary 2.29) or if the map $\mu \mapsto h_\mu(T)$ is upper semi-continuous for some other reason. The second equality in Theorem 2.22 holds by the ergodic decomposition of entropy in Theorem 4.23.

2.3.3 First Proof of Variational Principle

Proof of Theorem 2.22: topological entropy dominates. Let

$$\xi = \{P_1, \ldots, P_k\}$$

be a measurable partition of X, and fix a measure $\mu \in \mathscr{M}^T(X)$. Choose a positive ε with $\varepsilon < \frac{1}{k \log k}$. For each $P_j \in \xi$ there is a compact set $Q_j \subseteq P_j$ with $\mu(P_j \setminus Q_j) < \varepsilon$. Define a new partition $\eta = \{Q_0, Q_1, \ldots, Q_k\}$ where

$$Q_0 = X \setminus \bigcup_{j=1}^{k} Q_j.$$

Notice that $\mu(Q_0) < k\varepsilon$ and

$$\mu(Q_i \cap P_j) = \begin{cases} \mu(Q_i) \text{ if } i = j; \\ 0 \quad \text{if } 1 \leqslant i \neq j. \end{cases}$$

It follows that

$$\begin{aligned} H_\mu\left(\xi\middle|\eta\right) &= \sum_{i=0}^{k}\sum_{j=1}^{k} \mu(Q_i)\left[-\frac{\mu(Q_i \cap P_j)}{\mu(Q_i)} \log \frac{\mu(Q_i \cap P_j)}{\mu(Q_i)}\right] \\ &= \mu(Q_0) \underbrace{\sum_{j=1}^{k} -\frac{\mu(Q_0 \cap P_j)}{\mu(Q_0)} \log \frac{\mu(Q_0 \cap P_j)}{\mu(Q_0)}}_{\leqslant \log k} \\ &\leqslant \varepsilon k \log k < 1 \end{aligned} \tag{2.7}$$

by choice of ε. Notice that the cover number $N\left(\bigvee_{i=0}^{n-1} T^{-i}\eta\right)$ for a partition is simply the number of non-empty elements. Now $\mathscr{U} = \{Q_0 \cup Q_i \mid i = 1, \ldots, k\}$ is an open cover of X, and so by Proposition 1.5.

$$\begin{aligned} H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\eta\right) &\leqslant \log N\left(\bigvee_{i=0}^{n-1} T^{-i}\eta\right) \\ &\leqslant \log\left[N\left(\bigvee_{i=0}^{n-1} T^{-i}\mathscr{U}\right) \cdot 2^n\right] \end{aligned} \tag{2.8}$$

since every element of $\bigvee_{i=0}^{n-1} T^{-i}\mathscr{U}$ is the union of at most 2^n elements of $\bigvee_{i=0}^{n-1} T^{-i}\eta$. By Proposition 1.16(3),

$$h_\mu(T, \xi) \leqslant h_\mu(T, \eta) + H_\mu(\xi\big|\eta).$$

By the inequalities (2.7) and (2.8), it follows that

$$h_\mu(T) \leqslant h_{\text{top}}(T) + \log 2 + 1. \tag{2.9}$$

Now for any $n \geqslant 1$, $h_\mu(T^n) = nh_\mu(T)$ and $h_{\text{top}}(T^n) = nh_{\text{top}}(T)$ by Proposition 1.17(1) and Lemma 3.5, so equation (2.9) implies that

$$h_\mu(T) \leqslant h_{\text{top}}(T).$$

Since μ was arbitrary, this means that

$$\sup_{\mu \in \mathscr{M}^T(X)} h_\mu(T) \leqslant h_{\text{top}}(T). \tag{2.10}$$

□

Proof of Theorem 2.22: invariant measures with large entropy. Fix $\varepsilon > 0$ and let E_n be an (n, ε)-separated set with cardinality $s_n(\varepsilon)$. Define measures

$$\nu_n = \frac{1}{s_n} \sum_{x \in E_n} \delta_x$$

and

$$\mu_n = \frac{1}{n} \sum_{j=0}^{n-1} T_*^j \nu_n.$$

Choose a sequence (n_k) with $n_k \to \infty$ for which

$$\limsup_{n\to\infty} \frac{1}{n} \log s_n(\varepsilon) = \lim_{k\to\infty} \frac{1}{n_k} \log s_{n_k}(\varepsilon),$$

and

$$\lim_{k\to\infty} \mu_{n_k} = \mu \in \mathscr{M}^T(X).$$

This may be done by finding a sequence with the first property, and then choosing a subsequence of that sequence with the second property using the weak*-compactness property as in [38, Th. 4.1]. Let ξ be a measurable partition of X with $\mu(\partial P) = 0$ and $\operatorname{diam}(P) < \varepsilon$ for all $P \in \xi$ (such a partition always exists: start with an open cover by metric open balls $B_{\varepsilon_x}(x)$ for all $x \in X$, with $\varepsilon_x < \varepsilon/2$ chosen to have $\mu\left(\partial(B_{\varepsilon_x}(x))\right) = 0$ for all x, and use a finite subcover to define the partition). Then for x and y in the same atom $Q \in \bigvee_{i=0}^{n-1} T^{-i}\xi = \xi_n$,

$$\mathsf{d}(T^i x, T^i y) < \varepsilon$$

for $i = 0, \ldots, n-1$, so $|Q \cap E_n| \leqslant 1$. From the definition of ν_n, $\nu_n(Q)$ is either 0 or $\frac{1}{s_n(\varepsilon)}$. It follows that

$$\begin{aligned} H_{\nu_n}\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right) &= -\sum_Q \underbrace{\nu_n(Q)}_{0 \text{ or } 1/s_n(\varepsilon)} \log \nu_n(Q) \\ &= \log s_n(\varepsilon). \end{aligned} \tag{2.11}$$

However, to make use of this we need to start working with the (almost invariant) measure μ_n. Moreover, we will need to work with a fixed partition $\eta = \xi_m$ as $n \to \infty$ to make use of the weak*-convergence. To this end, fix an integer $m \geqslant 1$ and let $n = dm + r$. Then

$$\begin{aligned} H_{\mu_n}(\eta) &\geqslant \frac{1}{n} \sum_{j=0}^{n-1} H_{T_*^j \nu_n}(\eta) \\ &\geqslant \frac{1}{n} \sum_{k=0}^{m-1} \sum_{\ell=0}^{d-1} H_{\nu_n}\left(T^{-(\ell m+k)}\eta\right) \end{aligned} \tag{2.12}$$

by convexity of $t \mapsto \phi(t)$ (Lemma 1.4), the formula

$$H_{T_*^j\nu_n}(\eta) = H_{\nu_n}(T^{-j}\eta),$$

and by dropping the last r terms. Now for any k, $0 \leqslant k < m$, the partition

$$\bigvee_{i=0}^{n-1} T^{-i}\xi = \bigvee_{j=0}^{d-1} T^{-jm}\eta \vee \bigvee_{i=dm}^{n-1} T^{-i}\xi \tag{2.13}$$

is coarser than the partition

$$\bigvee_{i=0}^{k-1} T^{-i}\xi \vee \bigvee_{\ell=0}^{d-1} T^{-(\ell m+k)}\eta \vee \bigvee_{i=dm}^{n-1} T^{-i}\xi. \tag{2.14}$$

The relationship claimed between the partitions in equation (2.13) and in the expression (2.14) simply means that

$$[0, n-1] = \bigcup_{j=0}^{d-1} (jm + [0, m-1]) \cup [dm, n-1]$$

is a subset of

$$[0, k-1] \cup \bigcup_{\ell=0}^{d-1} (\ell m + k + [0, m-1]) \cup [dm, n-1].$$

It follows by the subadditivity of entropy that

$$H_{\nu_n}\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right) \leqslant \sum_{\ell=0}^{d-1} H_{\nu_n}\left(T^{-(\ell m+k)}\eta\right) + 2m\log|\xi|. \tag{2.15}$$

Thus

$$\begin{aligned}
H_{\mu_n}\left(\bigvee_{i=0}^{m-1} T^{-i}\xi\right) &\geqslant \frac{1}{n}\sum_{k=0}^{m-1}\sum_{\ell=0}^{d-1} H_{\nu_n}\left(T^{-(\ell m+k)}\eta\right) \quad \text{(by equation (2.12))} \\
&\geqslant \frac{1}{n}\sum_{k=0}^{m-1} H_{\nu_n}\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right) - \frac{2m^2}{n}\log|\xi| \\
&\qquad\qquad \text{(by equation (2.15))} \\
&= \frac{m}{n} H_\nu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right) - \frac{2m^2}{n}\log|\xi| \\
&= \frac{m}{n}\log s_n - \frac{2m^2}{n}|\xi| \quad \text{(by equation (2.11))}.
\end{aligned}$$

Using the sequence (n_k) with $n_k \to \infty$ as before to deduce that

$$H_\mu\left(\bigvee_{i=0}^{m-1} T^{-i}\xi\right) = \lim_{k\to\infty} H_{\mu_{n_k}}\left(\bigvee_{i=0}^{m-1} T^{-i}\xi\right) \tag{2.16}$$
$$\geqslant m \limsup_{n\to\infty} \frac{1}{n}\log s_n(\varepsilon)$$

where equation (2.16) holds by Lemma 2.8 since $\mu(\partial Q) = 0$ for $Q \in \eta$. Now let $m \to \infty$ to see that

$$h_\mu(T,\xi) = \lim_{m\to\infty} \frac{1}{m} H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right) \geqslant \limsup_{n\to\infty} \frac{1}{n}\log s_n(\varepsilon).$$

The measure μ potentially changes as ε changes, but nonetheless we can deduce that

$$\sup_{\mu\in\mathscr{M}^T(X)} h_\mu(T) \geqslant h_{\mathrm{top}}(T),$$

which with equation (2.10) proves the theorem. □

2.3.4 A Stronger Form of Variational Principle

The variational principle is a global statement about the collection of numbers

$$\{h_{\mathrm{top}}(T,\mathscr{U})\}$$

over all open covers $\mathscr{U}$ and the collection of numbers

$$\{h_\mu(T,\xi)\}$$

over all $\mu \in \mathscr{M}^T$ and partitions ξ, but says nothing about possible relationships between the topological entropy with respect to a specific cover and the measure-theoretic entropy with respect to a specific invariant measure and partition. In this section we present a result of Blanchard, Glasner and Host [10] (we follow their proof closely) which gives a local form of the variational principle in the following sense. Given an open cover $\mathscr{U}$, a Borel probability measure μ is constructed with the property that $h_\mu(T,\xi) \geqslant h_{\mathrm{top}}(T,\mathscr{U})$ for any partition ξ that *refines* U (that is, any partition ξ with the property that any atom of ξ is contained in an element of the cover $\mathscr{U}$).

The proof of the first inequality (2.10) in the variational principle is relatively straightforward. The result in this section gives a stronger local version of the more difficult reverse inequality, which constructs measures with entropy close to the topological entropy[(13)].

Theorem 2.23 (Blanchard, Glasner, Host). *Let $T : (X,\mathsf{d}) \to (X,\mathsf{d})$ be a continuous map on a compact metric space, and let $\mathscr{U} = \{U_1,\ldots,U_d\}$ be a finite open cover of X. Then there is a measure $\mu \in \mathscr{M}^T$ with the property that $h_\mu(T,\xi) \geqslant h_{\mathrm{top}}(T,\mathscr{U})$ for any partition ξ that refines $\mathscr{U}$.*

Corollary 2.24. *If there is a finite open over $\mathscr{U}$ with $h_{\mathrm{top}}(T,\mathscr{U}) = h_{\mathrm{top}}(T)$, then T has a maximal measure μ and a finite partition ξ with*

$$h_\mu(T) = h_\mu(T,\xi).$$

As in Proposition 2.21 and Theorem 2.22, the proof ends up by taking a weak*-limit of measures, but in contrast to the situation of Proposition 2.21, we have no set of distinguished points to start with. Instead a counting argument is used to find points whose orbit behavior under the map is sufficiently complex, and the invariant measure is produced as a limit of convex combinations of measures supported on these points.

For the counting argument, we start with a finite alphabet $A = |\mathscr{U}|$ of symbols, and write $|w| = n$ for the length of a word $w = w_1 \dots w_n \in A^n$ of n symbols. If u is a word of length $k \leqslant n$ and w is a word of length n, then write

$$\mathbf{d}_u(w) = \frac{1}{n-k+1} \left|\{i \mid 1 \leqslant i \leqslant n-k+1 \text{ and } w_i \dots w_{i+k-1} = u\}\right|$$

for the density of complete occurrences of u in w. Notice the two extreme cases: if u does not occur in w at all, then $\mathbf{d}_u(w) = 0$ and if (for example) $u = a$ and $w = a^n$ for a single symbol $a \in A$, then $\mathbf{d}_u(w) = 1$. We associate to a given $w \in A^n$ a probability measure (that is, a probability vector) on A^k by assigning the probability $\mathbf{d}_u(w)$ to the word u. We further define the k-entropy of the word w to be

$$H_k(w) = -\sum_{u \in A^k} \phi(\mathbf{d}_u(w)),$$

the entropy function applied to the probability vector $\big(\mathbf{d}_u(w) \mid u \in A^k\big)$.

To see why it is reasonable to view $H_k(w)$ as an entropy, notice that if w is a very long piece of a generic point in the full 2-shift with respect to the Bernoulli $(\frac{1}{2},\frac{1}{2})$-measure then we expect a word u of length $k \ll n$ to appear in w with frequency approximately $\frac{1}{2^k}$. Thus, roughly speaking, we expect that

$$\frac{1}{k} H_k(w) \approx -\frac{1}{k} \sum_{u \in A^k} \phi(2^{-k}) = -\frac{2^k}{k} \phi(2^{-k}) = \log 2$$

as $n \to \infty$. In fact the k-entropy is more closely related to the material on compression and coding from Section 1.5.

Lemma 2.25. *For any $h > 0$, $\varepsilon > 0$ and $k \geqslant 1$,*

$$|\{w \in A^n \mid H_k(w) \leqslant kh\}| \leqslant \mathrm{e}^{n(h+\varepsilon)}$$

for all sufficiently large n (depending on k, ε, and $|A|$).

Roughly speaking, Lemma 2.25 says that there cannot be too many words of length n whose complexity (as measured by the averaged k-entropy $\frac{1}{k} H_k(w)$) is bounded.

Proof of Lemma 2.25 for $k = 1$. Assume first that $k = 1$, so we are counting the appearance of single symbols in w. We may also think of the alphabet A as the set $\{1, 2, \ldots, |A|\}$. By counting words w of length n with q_i appearances of the symbol i, we see that

$$|\{w \in A^n \mid H_1(w) \leqslant h\}| = \sum_{\substack{\mathbf{q} \text{ with} \\ (2.18),(2.19)}} \frac{n!}{q_1! \cdots q_{|A|}!}, \tag{2.17}$$

where the sum is taken over those vectors $\mathbf{q} \in \mathbb{N}^{|A|}$ with

$$\sum_{i=1}^{|A|} q_i = n \tag{2.18}$$

and

$$-\sum_{i=1}^{|A|} \phi(q_i/n) \leqslant h. \tag{2.19}$$

By Stirling's theorem (see [40, Th. 8.6]) there are constants $C_1 \in (0,1), C_2 > 0$ with

$$C_1 (N/\mathrm{e})^N \sqrt{N} \leqslant N! \leqslant C_2 (N/\mathrm{e})^N \sqrt{N} \tag{2.20}$$

for all $N \geqslant 1$. We claim that there is some constant C depending on $|A|$ but not on n with

$$\frac{n!}{q_1! \cdots q_{|A|}!} \leqslant C\mathrm{e}^{-n\sum_{i=1}^{|A|} \phi(q_i/n)} \leqslant C\mathrm{e}^{nh}. \tag{2.21}$$

To see this, we treat each of the factors appearing in equation (2.20) in turn.

(1) The constant coefficients contribute a factor of $\frac{C_2}{C_1^{|A|}}$.

(2) The term N^N contributes a factor

$$\begin{aligned}\frac{n^n}{\prod_{q_i>0} q_i^{q_i}} &= \exp\left(n \log n - \sum_{q_i>0} q_i \log q_i\right) \\ &= \exp\left(\sum_{q_i>0} q_i(\log n - \log q_i)\right) = \exp\left(-n \sum_{q_i>0} \frac{q_i}{n} \log\left(\frac{q_i}{n}\right)\right)\end{aligned}$$

by equation (2.18), giving $\mathrm{e}^{-n\sum_{i=1}^{|A|} \phi(q_i/n)}$ since $\phi(0) = 0$.

(3) The term e^{-N} contributes a factor $\frac{\mathrm{e}^{-n}}{\mathrm{e}^{-q_1} \cdots \mathrm{e}^{-q_{|A|}}} = 1$ by equation (2.18).

(4) For the term $\sqrt{N}$, notice that

$$n = \sum_{q_i>0} q_i \leqslant |A| \max_{q_i>0} q_i \leqslant |A| \prod_{q_i>0} q_i,$$

giving a factor $\sqrt{|A|}$.

This gives the bound (2.21), with $C = \frac{C_2}{C_1^{|A|}}\sqrt{|A|}$.

The number of terms summed in equation (2.17) is no more than

$$(n+1)^{|A|},$$

since each symbol in A appears no more than n times, so

$$\begin{aligned} |\{w \in A^n \mid H_1(w) \leqslant h\}| &\leqslant C(n+1)^{|A|}\mathrm{e}^{nh} \qquad \text{(by (2.21))} \\ &\leqslant \mathrm{e}^{n(h+\varepsilon)}, \end{aligned}$$

for sufficiently large n, for fixed $\varepsilon > 0$, as required. □

As we will see, if $\frac{1}{k}H_k(w)$ is small, then it is possible to group the word w into blocks of length k (possibly using an offset j) so that the new word $w^{(j)}$ consisting of roughly $\frac{n}{k}$ blocks of length k has a small value for $H_1(w^{(j)})$, which will allow us to use the case considered above. This may be seen at equation (2.22) below.

Proof of Lemma 2.25 for $k > 1$. Now assume that $n > 2k$ and $k > 1$. We will think of the word w on two different scales. It is a sequence of symbols of length n; on the other hand it is also (up to remainder terms at the ends) a sequence of $m = \lfloor \frac{n}{k} \rfloor - 1$ words of length k.

For each j, $0 \leqslant j < k$, we define the word $w^{(j)}$ to be the word comprising m words of length k defined by

$$w^{(j)} = |w_{j+1} \dots w_{j+k}|w_{j+k+1} \dots w_{j+2k}| \cdots |w_{j+m(k-1)+1} \dots w_{j+mk}|,$$

viewed as a sequence of length m on the alphabet A^k. Here j is a shift determining which initial and final symbols are discarded when switching from w to $w^{(j)}$.

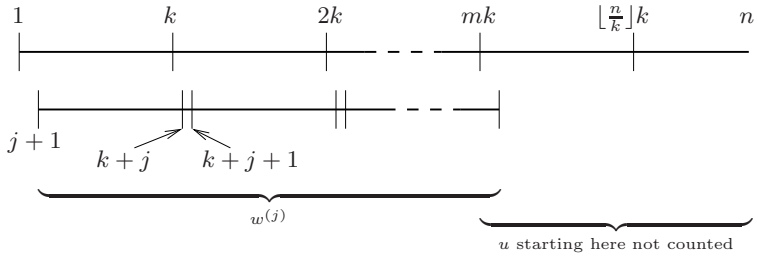

Fig. 2.1. Counting blocks in $w^{(j)}$.

Now if $I_u(w)$ denotes the number of positions where an incidence of u starts in w, so that

$$I_u(w) = (n-k+1)\mathbf{d}_u(w),$$

then we claim that

$$I_u(w) - 2k \leqslant \sum_{j=0}^{k-1} \mathbf{d}_u(w^{(j)})m \leqslant I_u(w).$$

To see the upper bound, notice that every occurrence of u in $w^{(j)}$ is an occurrence in w at some offset j; for the lower bound, notice that occurrences of u near the end of w may be missed. Now $\frac{m}{n} \to \frac{1}{k}$ as $n \to \infty$, so it follows that

$$\left| \mathbf{d}_u(w) - \frac{1}{k} \sum_{j=0}^{k-1} \mathbf{d}_u(w^{(j)}) \right| = \mathrm{O}\left(\tfrac{1}{n}\right)$$

as $n \to \infty$. Since ϕ is uniformly continuous, it follows that for large enough n (depending on k and ε) and for any word w of length n,

$$\sum_{u \in A^k} \left| \phi(\mathbf{d}_u(w)) - \phi\left(\frac{1}{k} \sum_{j=0}^{k-1} \mathbf{d}_u(w^{(j)}) \right) \right| < \frac{\varepsilon}{2}.$$

By convexity of ϕ (Lemma 1.4), this gives

$$\begin{aligned} \frac{1}{k} \sum_{j=0}^{k-1} H_1\left(w^{(j)}\right) &= - \sum_{u \in A^k} \frac{1}{k} \sum_{j=0}^{k-1} \phi\left(\mathbf{d}_u(w^{(j)})\right) \\ &\leqslant - \sum_{u \in A^k} \phi\left(\sum_{j=0}^{k-1} \frac{1}{k} \mathbf{d}_u(w^{(j)}) \right) \\ &\leqslant \frac{\varepsilon}{2} - \sum_{u \in A^k} \phi\left(\mathbf{d}_u(w)\right) = \frac{\varepsilon}{2} + H_k(w). \end{aligned} \tag{2.22}$$

It follows that if $H_k(w) \leqslant kh$ then there is some j with $H_1(w^{(j)}) \leqslant \frac{\varepsilon}{2} + kh$. Now for any $j \geqslant 1$ and word $\widetilde{w}$ of length m in the alphabet A^k, there are at most

$$|A|^{n-mk} \leqslant |A|^{2k}$$

words w of length n in the alphabet A for which $w^{(j)} = \widetilde{w}$ (see Figure 2.1). It follows from the case $k = 1$ that

$$\begin{aligned} |\{w \in A^n \mid H_k(w) \leqslant kh\}| &\leqslant |A|^{2k} \sum_{j=0}^{k-1} \left| \left\{ \widetilde{w} \in \left(A^k\right)^m \mid H_1(\widetilde{w}) \leqslant \frac{\varepsilon}{2} + kh \right\} \right| \\ &\leqslant |A|^{2k} \sum_{j=0}^{k-1} \mathrm{e}^{m(\varepsilon + kh)} \end{aligned}$$

for sufficiently large n. We deduce that

$$|\{w \in A^n \mid H_k(w) \leqslant kh\}| \leqslant k|A|^{2k} \mathrm{e}^{n(\varepsilon/k+h)} \leqslant \mathrm{e}^{n(h+\varepsilon)}$$

for large n. □

In order to connect the behavior of partitions to the combinatorics of words, we will use the *names* associated to a partition (this idea will be used extensively in Chapter 6). A finite partition $\xi = \{P_1, \dots, P_d\}$ defines, for each $N \geqslant 1$, a map

$$\mathbf{w}_N^\xi : X \to \{1, \dots, d\}^N$$

by requiring that the nth coordinate of $\mathbf{w}_N^\xi(x)$ is j if $T^{n-1}x \in P_j$, $1 \leqslant n \leqslant N$.

We will also reduce Theorem 2.23 to the case of a zero-dimensional dynamical system, that is a continuous map on a zero-dimensional compact metric space*.

Lemma 2.26. *Any topological dynamical system is a topological factor of a zero-dimensional dynamical system.*

Proof. Let $T : X \to X$ be a continuous map on a compact metric space; the claim of the lemma is that we can find a continuous map $S : Y \to Y$ of a zero-dimensional space together with a continuous surjective map $\pi : Y \to X$ with $\pi \circ S = T \circ \pi$.

Since X is compact, we may find a sequence of finite open covers

$$\mathscr{U}_1 \leqslant \mathscr{U}_2 \leqslant \cdots$$

with $\operatorname{diam}(\mathscr{U}_n) \to 0$ as $n \to \infty$. Let a_n be the number of elements in $\mathscr{U}_n$, and fix an enumeration $\mathscr{U}_n = \{O_n^1, \dots, O_n^{a_n}\}$. Define a map sending an element $(x_1, x_2, \dots) \in K_0 = \prod_{n \geqslant 1} \{1, 2, \dots, a_n\}$ (which is a compact metric space in the product topology) to the unique element $x \in X$ with the property that x belongs to the closure $\overline{O_n^{x_n}}$ of the x_nth element of $\mathscr{U}_n$ for all $n \geqslant 1$, if there is such an element. By a straightforward compactness argument, this procedure defines a surjective map from a closed subset $K \subseteq K_0$ onto X.

Now let

$$Y = \{y \in K^{\mathbb{Z}} \mid \theta(y_{n+1}) = T(\theta(y_n)) \text{ for all } n \in \mathbb{Z}\},$$

and define $\pi : Y \to X$ by $\pi(y) = \theta(y_0)$. Since T and θ are continuous, Y is a closed subset of the compact set $K^{\mathbb{Z}}$ with the product topology, and it is clear that π is continuous and onto. Finally, if we write $S : Y \to Y$ for the left shift map defined by $(S(y))_k = y_{k+1}$ for all $k \in \mathbb{Z}$, we have $\pi \circ S = T \circ \pi$. □

* A topological space is said to be zero-dimensional if there is a basis for the topology comprising sets that are both open and closed (clopen sets). Any discrete space is zero-dimensional, but a zero-dimensional space need not have any isolated points. Examples include $\mathbb{Q}$ in the subspace topology induced from the reals. The examples *ne plus ultra* for dynamics come from shift spaces: for any finite set A with the discrete topology, the spaces $A^{\mathbb{Z}}$, $A^{\mathbb{N}}$, or any closed subset of them, are zero-dimensional.

Proof of Theorem 2.23: reduction to zero-dimensional case. Assume first that we have proved the theorem for any zero-dimensional topological dynamical system, and let $\pi : Y \to X$ be a topological factor map from a zero-dimensional system $S : Y \to Y$, which exists by Lemma 2.26. Let $\mathscr{V} = \pi^{-1}(\mathscr{U})$ be the pre-image of $\mathscr{U}$, so that $h_{\mathrm{top}}(S, \mathscr{V}) = h_{\mathrm{top}}(T, \mathscr{U})$. By Theorem 2.23 for zero-dimensional systems, there is a measure $\nu \in \mathscr{M}^S(Y)$ with the property that $h_\nu(S, \eta) \geqslant h_{\mathrm{top}}(S, \mathscr{V})$ for every measurable partition η finer than $\mathscr{V}$. Let $\mu = \pi_*\nu$; then $\mu \in \mathscr{M}^T(X)$, and for any measurable partition ξ finer than $\mathscr{U}$, $\pi^{-1}(\xi)$ is a measurable partition of Y that refines $\mathscr{V}$, so

$$h_\mu(T, \xi) = h_\nu(S, \pi^{-1}(\xi)) \geqslant h_{\mathrm{top}}(S, \mathscr{V}) = h_{\mathrm{top}}(T, \mathscr{U})$$

as required. □

All that remains is to prove the theorem for a zero-dimensional system. We start with a lemma which relates partitions to names and the combinatorial entropy of blocks in those names.

Lemma 2.27. *If $\mathscr{U}$ is a finite cover of X, $K \in \mathbb{N}$ and $\{\xi_\ell \mid 1 \leqslant \ell \leqslant K\}$ is a finite list of finite measurable partitions of X each of which refines $\mathscr{U}$, then for any $\varepsilon > 0$ and sufficiently large n, there is an $x \in X$ with*

$$H_k\left(\mathbf{w}_N^{\xi_\ell}(x)\right) \geqslant k\left(h_{\mathrm{top}}(T, \mathscr{U}) - \varepsilon\right)$$

if $1 \leqslant k, \ell \leqslant K$.

Proof of Lemma 2.27. By allowing empty sets and taking unions of atoms that lie in the same element of $\mathscr{U}$, we may assume that all the partitions ξ_ℓ have $d = |\mathscr{U}|$ elements; write $A = \{1, \ldots, d\}$. Then, if n is large enough, we have by Lemma 2.25

$$|\Sigma(n, k)| \leqslant \mathrm{e}^{n(h_{\mathrm{top}}(T, \mathscr{U}) - \varepsilon/2)}$$

for $1 \leqslant k \leqslant K$, where

$$\Sigma(n, k) = \{w \in A^n \mid H_k(w) < k\left(h_{\mathrm{top}}(T, \mathscr{U}) - \varepsilon\right)\}.$$

By further increasing n if necessary, also assume that

$$\mathrm{e}^{n\varepsilon/2} > K^2.$$

Write

$$\Upsilon(k, \ell) = \{x \in X \mid \mathbf{w}_n^{\xi_\ell}(x) \in \Sigma(n, k)\};$$

since $\Upsilon(k, \ell)$ is the union of $|\Sigma(n, k)|$ elements of $\bigvee_{j=0}^{n-1} T^{-j}\xi_\ell$, a partition finer than $\bigvee_{j=0}^{n-1} T^{-j}\mathscr{U}$, the set $\Upsilon(k, \ell)$ is covered by

$$|\Sigma(n, k)| \leqslant \mathrm{e}^{n(h_{\mathrm{top}}(T, \mathscr{U}) - \varepsilon/2)}$$

elements of $\bigvee_{j=0}^{n-1} T^{-j}\mathscr{U}$. It follows that $\bigcup_{1\leqslant k,\ell\leqslant K}\Upsilon(k,\ell)$ is covered by

$$K^2\mathrm{e}^{n(h_{\mathrm{top}}(T,\mathscr{U})-\varepsilon/2)} < \mathrm{e}^{nh_{\mathrm{top}}(T,\mathscr{U})}$$

elements of $\bigvee_{j=0}^{n-1} T^{-j}\mathscr{U}$. Since

$$h_{\mathrm{top}}(T,\mathscr{U}) \leqslant \frac{1}{n}\log N\left(\bigvee_{j=0}^{n-1} T^{-j}\mathscr{U}\right),$$

by Fekete's lemma (Lemma 1.13, as used in Definition 2.10), any subcover of

$$\bigvee_{j=0}^{n-1} T^{-j}\mathscr{U}$$

has at least $\mathrm{e}^{nh_{\mathrm{top}}(T,\mathscr{U})}$ elements, so we deduce that

$$\bigcup_{1\leqslant k,\ell\leqslant K}\Upsilon(k,\ell) \neq X,$$

and the point x may be found in the complement. □

Proof of Theorem 2.23: zero-dimensional case. Assume now that X is zero-dimensional and let $\mathscr{U} = \{U_1,\ldots,U_d\}$ be the open cover of X. Consider initially the collection Ξ of all partitions ξ of X with the property that

$$\xi = \{P_1,\ldots,P_d\}$$

comprises d clopen sets with $P_i \subseteq U_i$ for all $i = 1,\ldots,d$ (we will see later how to extend the result to all partitions). The collection Ξ is countable*, so we may enumerate it as $\Xi = \{\xi_\ell \mid \ell \geqslant 1\}$. Using Lemma 2.27, we may find a sequence of integers $n_K \to \infty$ and a sequence (x_K) in X for which

$$H_k\left(\mathbf{w}_{n_K}^{\xi_\ell}(x_K)\right) \geqslant k\left(h_{\mathrm{top}}(T,\mathscr{U}) - \tfrac{1}{K}\right) \tag{2.23}$$

for $1 \leqslant k,\ell \leqslant K$. Define a measure μ_K by

$$\mu_K = \frac{1}{n_K}\sum_{i=0}^{n_K-1}\delta_{T^i x_K},$$

and (by passing to a subsequence of (n_K) and using the corresponding subsequence of (x_K) with the property (2.23); for brevity we use K again to index

* To see this, notice that for any $\varepsilon = \frac{1}{n}$ the whole space can be covered by finitely many clopen sets of diameter less than ε. Denote by Ξ_0 the countable collection of finite unions of such sets. Then Ξ_0 is the collection of all clopen sets: any open set is a union of elements of Ξ_0 and a compact open set is a finite union.

the resulting sequences) we can assume that $\mu_K \to \mu$ in the weak*-topology. By [38, Th. 4.1], $\mu \in \mathscr{M}^T(X)$. Fix k and ℓ and let E be any atom of the partition $\bigvee_{j=0}^{k-1} T^{-j}\xi_\ell$ with name $u \in \{1,\dots,d\}^k$. For every K,

$$\left|\mu_K(E) - \mathbf{d}_u(\mathbf{w}_{n_K}^{\xi_\ell}(x_K))\right| = \mathrm{O}\left(\tfrac{k}{n_K}\right),$$

and, since E is clopen and therefore χ_E is continuous,

$$\mu(E) = \lim_{K\to\infty} \mu_K(E) = \lim_{K\to\infty} \mathbf{d}_u(\mathbf{w}_{n_K}^{\xi_\ell}(x_K)),$$

so

$$\phi\left(\mu(E)\right) = \lim_{K\to\infty} \phi\left(\mathbf{d}_u(\mathbf{w}_{n_K}^{\xi_\ell}(x_K)\right). \tag{2.24}$$

Summing equation (2.24) over all $u \in \{1,\dots,d\}^k$ and using equation (2.23) gives

$$H_\mu\left(\bigvee_{j=0}^{k-1} T^{-j}\xi_\ell\right) = \lim_{K\to\infty} H_k\left(\mathbf{w}_{n_K}^{\xi_\ell}(x_K)\right) \geqslant k h_{\mathrm{top}}(T,\mathscr{U}),$$

and letting $k \to \infty$ then gives $h_\mu(T,\xi_\ell) \geqslant h_{\mathrm{top}}(T,\mathscr{U})$ for all $\ell \geqslant 1$.

Finally, since X is zero-dimensional the family Ξ of partitions using clopen sets is dense (with respect to the L^1_μ metric on partitions) in the family of partitions with d atoms each of which is a subset of an atom of $\mathscr{U}$. Together with the continuity of entropy from Proposition 1.16 we also have

$$h_\mu(T,\xi) \geqslant h_{\mathrm{top}}(T,\mathscr{U})$$

for any partition of this shape, proving the theorem. □

Exercises for Section 2.3

Exercise 2.3.1. Let $T_p(x) = px \mod 1$ for $x \in \mathbb{T}$ for some $p \geqslant 2$, and assume that for every n with $\gcd(n,p) =$ we choose a subset $S_n \subseteq \{0, \frac{1}{n}, \dots, \frac{n-1}{n}\}$ with $T_p(S_n) \subseteq S_n$ and $|S_n| \geqslant n^{1-\mathrm{o}(1)}$ (that is, there is a sequence $a_n \to 0$ with $|S_n| \geqslant n^{1-a_n}$). Use Proposition 2.21 to show that the sequence of sets (S_n) is equidistributed: for any continuous function $f : \mathbb{T} \to \mathbb{R}$,

$$\frac{1}{|S_n|}\sum_{s\in S_n} f(s) \longrightarrow \int f \,\mathrm{d}m_{\mathbb{T}}$$

as $n \to \infty$.

Exercise 2.3.2. Fill in the details of the compactness argument in the proof of Lemma 2.26.

2.4 Maximal Measures

As mentioned above, a measure $\mu \in \mathscr{M}^T(X)$ with $h_\mu(T) = h_{\text{top}}(T)$ is called a *maximal measure*. It is clear from Theorem 4.23 that the set of maximal measures is a (possibly empty) convex subset of $\mathscr{M}^T$, and indeed it shares other properties with $\mathscr{M}^{T(14)}$.

Lemma 2.28. *Let $T : X \to X$ be a continuous map on a compact metric space. If $h_{\text{top}}(T) < \infty$ and T has a measure of maximal entropy, then it has an ergodic measure of maximal entropy.*

Proof. Let μ be a maximal measure; using [38, Th. 6.2] write

$$\mu = \int_Y \mu_y \, \mathrm{d}\nu(y)$$

for the ergodic decomposition of μ. By Theorem 4.23,

$$h_{\text{top}}(T) = h_\mu(T) = \int_Y h_{\mu_y}(T) \, \mathrm{d}\nu(y).$$

Since $h_{\mu_y}(T) \leqslant h_{\text{top}}(T)$ for each $y \in Y$, we must have

$$h_{\mu_y}(T) = h_\mu(T) = h_{\text{top}}(T)$$

for ν-almost every y, and in particular there must be an ergodic measure of maximal entropy. □

The next result records one general situation in which at least one maximal measure exists[15]; it is an immediate corollary of Theorem 2.7.

Corollary 2.29. *If $T : (X, \mathsf{d}) \to (X, \mathsf{d})$ is an expansive homeomorphism, then T has a maximal measure.*

Proof. By Theorem 2.7, the entropy map $\mu \mapsto h_\mu(T)$ is upper semi-continuous, and an upper semi-continuous real-valued map on a compact space attains its bounds. □

The simplest[16] example of a unique measure of maximal entropy arises for full shifts.

Lemma 2.30. *Let $X = \prod_{-\infty}^{\infty} \{0, 1, \ldots, s-1\}$ with the metric*

$$\mathsf{d}(x, y) = \sum_{n \in \mathbb{Z}} |x_i - y_i| \cdot 2^{-|n|};$$

the shift map $\sigma : X \to X$ defined by $(\sigma(x))_k = x_{k+1}$ for all $k \in \mathbb{Z}$ is a homeomorphism of (X, d). The map σ has a unique measure of maximal entropy, and this unique measure is the Bernoulli measure corresponding to the uniform measure $(\frac{1}{s}, \ldots, \frac{1}{s})$ on the alphabet $\{0, 1, \ldots, s-1\}$.

PROOF. Notice that Corollary 2.14 applies equally well to two-sided shifts (see Exercise 2.2.1), so $h_{\mathrm{top}}(\sigma) = \log s$. Similarly, the argument from Example 1.25 shows that the measure-theoretic entropy with respect to the uniform $(\frac{1}{s}, \dots, \frac{1}{s})$ measure is $\log s$.

Now let $\mu \in \mathscr{M}^T(X)$ be any measure with $h_\mu(T) = \log s$, and let

$$\xi = \{\{x \in X \mid x_0 = j\} \mid j = 0, \dots, s-1\}$$

be the state partition. Then by Lemma 1.13 and Proposition 1.5,

$$\log s = h_\mu(T) = \inf_{n \geqslant 1} \frac{1}{n} H_\mu \left(\bigvee_{i=0}^{n-1} T^{-i}\xi \right) \leqslant \frac{1}{m} \log s^m = \log s$$

for any $m \geqslant 1$. By Proposition 1.5 again, it follows that each atom in the partition $\bigvee_{i=0}^{m-1} T^{-i}\xi$ must have measure $\frac{1}{s^m}$, showing that μ must be the measure claimed. □

Two more examples of existence and uniqueness of maximal measures, namely the circle doubling map and the toral automorphism $(x, y) \mapsto (y, x+y)$ on $\mathbb{T}^2$, have been discussed in Section 4.5.3[(17)].

2.4.1 Examples Without Maximal Measures

Next we record a simple example to show that there may be no maximal measures at all. The first such examples were found by Gurevič [57].

Example 2.31. For each $N \geqslant 2$, define a closed σ-invariant subset of the full shift $X = \prod_{-\infty}^{\infty} \{0, 1\}$ by

$$X_{(N)} = \{x \in X \mid \text{the block } 0^N \text{ does not appear in } x\}$$

and write $\sigma_{(N)}$ for the shift restricted to $X_{(N)}$. For any $k \geqslant 1$ there are no more than $(2^N - 1)^k$ blocks of length Nk, because there are $(2^N - 1)$ allowed blocks of length N and not all of them can be concatenated to make an allowed block. On the other hand, if $w_1, \dots, w_k$ are arbitrary blocks of length $(N-1)$ then $w_1 1 w_2 1 \cdots 1 w_k$ is an allowed block of length $Nk - 1$, so there are at least $(2^{N-1})^k$ blocks of length $Nk - 1$. By the two-sided version of Corollary 2.14 and Lemma 3.5 we deduce that

$$\left(1 - \tfrac{1}{N}\right) \log 2 \leqslant h_{\mathrm{top}}(\sigma_{(N)}) \leqslant \tfrac{1}{N} \log(2^N - 1),$$

and, in particular, $h_{\mathrm{top}}(\sigma_{(N)}) < \log 2$ while $h_{\mathrm{top}}(\sigma_{(N)}) \to \log 2$ as $N \to \infty$.

Choose a metric d_N on each $X_{(N)}$ compatible with its compact topology and with $\mathrm{diam}_{\mathsf{d}_N}(X_{(N)}) = 1$. Define a new space X_* to be the disjoint union

$$X = \bigsqcup_{N \geqslant 1} X_{(N)} \sqcup \{\infty\}$$

of all the sets $X_{(N)}$ with an additional point at infinity. Make X_* into a metric space by defining

$$\mathsf{d}(x,y) = \begin{cases} \dfrac{1}{N^2}\mathsf{d}_N(x,y) & \text{if } x,y \in X_{(N)}; \\ \displaystyle\sum_{j=M}^{N} \frac{1}{j^2} & \text{if } x \in X_{(M)}, y \in X_{(N)} \text{ and } M < N; \\ \displaystyle\sum_{j=M}^{\infty} \frac{1}{j^2} & \text{if } x \in X_{(M)} \text{ and } y = \infty, \end{cases}$$

and then extending d to a metric by requiring that $\mathsf{d}(x,y) = \mathsf{d}(y,x)$ for all $x, y \in X_*$. The space (X_*, d) is compact, and

$$\sigma_*(x) = \begin{cases} \sigma_{(N)}(x) & \text{if } x \in X_{(N)}; \\ \infty & \text{if } x = \infty \end{cases}$$

defines a homeomorphism σ_* of X_*. Now let μ be any σ_*-invariant probability measure on X_* and allow N to denote a member of $\mathbb{N}$ or ∞. Since X_* is a disjoint union, we may write $\mu = \sum_{N \leqslant \infty} p_N \mu_N$, where $p_N \in [0,1]$ for all $N \leqslant \infty$, $\sum_{N \leqslant \infty} p_N = 1$, $\mu_N \in \mathscr{M}^{\sigma_{(N)}}(X_{(N)})$ for $N \geqslant 1$, and $\mu_\infty = \delta_\infty$. If μ is ergodic then by [38, Th. 4.4], $\mu \in \mathscr{E}^{\sigma_{(N)}}(X_{(N)})$ for some $N \geqslant 1$, or $\mu = \mu_\infty$. By Exercise 4.4.3,

$$\begin{aligned} h_{\mathrm{top}}(\sigma_*) &= \sup\{h_\mu(\sigma_*) \mid \mu \in \mathscr{E}^{\sigma_*}(X_*)\} \\ &= \sup\{h_{\mu_N}(\sigma_{(N)}) \mid \mu_N \in \mathscr{E}^{\sigma_{(N)}}(X_{(N)})\} \quad (\text{since } h_{\mu_\infty}(\sigma) = 0) \\ &= \sup\{h_{\mathrm{top}}(\sigma_{(N)}) \mid N \geqslant 1\} = \log 2. \end{aligned}$$

If σ has a measure of maximal entropy μ, then by Lemma 2.28 there is an ergodic measure of maximal entropy, so $h_\mu(\sigma) = h_{\mu_N}(\sigma_{(N)})$ for some N. On the other hand, by construction μ_N cannot be the maximal measure for the full shift, so $h_{\mu_N}(\sigma_{(N)}) < \log 2$, which contradicts maximality. Thus (X_*, σ_*) has no measure of maximal entropy.

Notes to Chapter 2

[10](Page 38) An alternative and equally natural definition is the following: A homeomorphism $T : (X, \mathsf{d}) \to (X, \mathsf{d})$ of a compact metric space is called *pointwise expansive* if there is a map $\delta : X \to \mathbb{R}_{>0}$ such that $\mathsf{d}(T^n x, T^n y) \leqslant \delta(x)$ for all $n \in \mathbb{Z}$ implies $x = y$. This shares many of the properties of expansiveness. For example if T is pointwise expansive then T has finitely many fixed points and at most countably many periodic points; there are no pointwise expansive homeomorphisms of a circle. These observations are due to Reddy [120], who also showed that there are pointwise expansive homeomorphisms that are not expansive.

[11](Page 39) The definition of expansiveness does not require the space to be compact, and the local isometry $\mathbb{R}^k \to \mathbb{R}^k/\mathbb{Z}^k \cong \mathbb{T}^k$ may be used to show that the

toral automorphism corresponding to $A \in \mathrm{GL}_k(\mathbb{Z})$ is expansive if and only if the action of A on $\mathbb{R}^k$ is expansive. Studying the geometry of the action of the Jordan form of the complexification of A gives the result (see Eisenberg [39]). A similar argument will be used in Section 3.3 to compute the topological entropy of toral automorphisms.

(12) (Page 49) The inequality $h_\mu(T) \leqslant h_{\mathrm{top}}(T)$ for all $\mu \in \mathscr{M}^T(X)$ was shown by Goodwyn [55]; Dinaburg [31] then proved Theorem 2.22 under the assumption that $\dim(X) < \infty$, and finally Goodman [53] proved the general case. The variational principle is generalized significantly by Walters [139] using the notion of topological pressure. It is extended in a different direction by Blanchard, Glasner and Host [10], and their result will be described in Section 2.3.4.

(13) (Page 55) This result is also presented in Glasner's monograph [49]. A similar local strengthening of the reverse inequality, namely the result that

$$\sup_{\mu \in \mathscr{M}^T} \inf_{\xi \succcurlyeq U} h_\mu(T, \xi) = h_{\mathrm{top}}(T, U)$$

(where the infimum is taken over all partitions that refine the open cover U) was found by Glasner and Weiss [51]. In [10, Prop. 4] an example is given of a dynamical system (X, T) and an open cover U of X with the property that $h_{\mathrm{top}}(T, U) > 0$, but for every ergodic measure $\mu \in \mathscr{M}^T$ we have $h_\mu(T, \xi) = 0$ for some partition ξ that refines U. There are subsequent developments, many of which may be found in a review paper of Glasner and Ye [52].

(14) (Page 64) For example, the extreme points are ergodic measures when T has finite topological entropy. See Walters [140, Sect. 8.3] or Denker, Grillenberger and Sigmund [30] for a detailed treatment.

(15) (Page 64) Clearly a uniquely ergodic map has a maximal measure; what is much less clear is whether this arises in an interesting way. The examples we have seen of uniquely ergodic maps seem to have zero entropy. A deep result of Hahn and Katznelson [58] is that there are minimal uniquely ergodic maps with positive entropy. Corollary 2.29 was shown by Dinaburg [32] under the assumption that $\dim(X) < \infty$. Corollary 2.29 as stated is taken from Goodman [54].

(16) (Page 64) This is a special case of a more general result due to Parry [111], which shows that any topologically mixing subshift of the full shift defined by forbidding a finite list of blocks has a unique measure of maximal entropy.

(17) (Page 65) In light of Corollary 2.29, a natural question is to ask if an expansive homeomorphism has a unique maximal measure. It is clear that this is not the case if the map is not minimal, so the question is of interest for minimal expansive homeomorphisms, and Goodman [54] gives examples of minimal expansive maps of positive entropy with more than one maximal measure. In contrast to the case of a topologically mixing one-dimensional shift obtained by forbidding finitely many blocks (see note (16)), Burton and Steif [20] showed that there are topologically mixing two-dimensional shifts obtained by forbidding finitely many blocks with many maximal measures for the resulting $\mathbb{Z}^2$ shift action.

3

Lifting Entropy

In this chapter we will extend the theory of topological entropy to uniformly continuous maps on metric spaces, and use this method (introduced by Bowen [14]) to compute the topological entropy of automorphisms of solenoids including the torus.

3.1 Entropy for Uniformly Continuous Maps

Suppose that (X, d) is a locally compact σ-compact metric space, and that

$$T : X \to X$$

is a uniformly continuous map (that is, for any $\varepsilon > 0$ there is a $\delta > 0$ such that $\mathsf{d}(x, y) < \delta$ implies $\mathsf{d}(f(x), f(y)) < \varepsilon$ for all $x, y \in X$). Definition 2.16 extends to this setting as follows. For a compact set $K \subseteq X$, we say that a subset $F \subseteq K$ (n, ε)-spans K if, for every $x \in K$ there is a point $y \in F$ with $\mathsf{d}(T^i x, T^i y) \leqslant \varepsilon$ for $i = 0, \ldots, n-1$ and a subset $E \subseteq K$ is (n, ε)-separated if for any two distinct points $x, y \in E$,

$$\max_{0 \leqslant i \leqslant n} \mathsf{d}(T^i x, T^i y) > \varepsilon.$$

Let $r_n(\varepsilon, K, \mathsf{d})$ denote the smallest cardinality of any set F which (n, ε)-spans K with respect to T, and let $s_n(\varepsilon, K, \mathsf{d})$ denote the largest cardinality of any (n, ε)-separated set with respect to T contained in K. As before, the compactness of K ensures that $s_n(\varepsilon, K, \mathsf{d})$ is finite. Just as in Lemma 2.17, we have that

$$r_n(\varepsilon, K, \mathsf{d}) \leqslant s_n(\varepsilon, K, \mathsf{d}) \leqslant r_n(\varepsilon/2, K, \mathsf{d}) < \infty; \tag{3.1}$$

it is also clear that if $\varepsilon < \varepsilon'$ then

$$\limsup_{n\to\infty} \frac{1}{n} \log r_n(\varepsilon, K, \mathsf{d}) \geqslant \limsup_{n\to\infty} \frac{1}{n} \log r_n(\varepsilon', K, \mathsf{d})$$

and

$$\limsup_{n\to\infty}\frac{1}{n}\log s_n(\varepsilon,K,\mathsf{d})\geqslant\limsup_{n\to\infty}\frac{1}{n}\log s_n(\varepsilon',K,\mathsf{d}). \tag{3.2}$$

An immediate consequence of equations (3.1)–(3.2) is that the following definition makes sense.

Definition 3.1. *The Bowen entropy of T with respect to K is*

$$h_{\mathsf{d}}(T,K)=\lim_{\varepsilon\searrow 0}\limsup_{n\to\infty}\frac{1}{n}\log r_n(\varepsilon,K,\mathsf{d})=\lim_{\varepsilon\searrow 0}\limsup_{n\to\infty}\frac{1}{n}\log s_n(\varepsilon,K,\mathsf{d}),$$

and the Bowen entropy of T is

$$h_{\mathsf{d}}(T)=\sup_{K\subseteq X\ compact}h_{\mathsf{d}}(T,K).$$

The notation reflects the fact that the quantities all depend[(18)] on the choice of the metric d. Metrics d' and d on X are called *uniformly equivalent* if for any $\varepsilon>0$ there are constants $\varepsilon',\varepsilon''>0$ for which

$$\mathsf{d}(x,y)<\varepsilon'\implies\mathsf{d}'(x,y)<\varepsilon,$$

and

$$\mathsf{d}''(x,y)<\varepsilon\implies\mathsf{d}'(x,y)<\varepsilon$$

for all $x,y\in X$.

Lemma 3.2. *If d' is uniformly equivalent to d, then $h_{\mathsf{d}}(T)=h_{\mathsf{d}'}(T)$.*

Proof. Fix $\varepsilon>0$ and a compact set $K\subseteq X$. By assumption, there is a constant $\varepsilon'>0$ with the property that $\mathsf{d}'(x,y)\leqslant\varepsilon'\implies\mathsf{d}(x,y)\leqslant\varepsilon$ for all $x,y\in X$. An (n,ε')-spanning set for K with respect to d' is an (n,ε)-spanning set for K with respect to d, so $r_n(\varepsilon,K,\mathsf{d})\leqslant r_n(\varepsilon',K,\mathsf{d}')$. It follows that

$$\limsup_{n\to\infty}\frac{1}{n}\log r_n(\varepsilon,K,\mathsf{d})\leqslant\limsup_{n\to\infty}\frac{1}{n}\log r_n(\varepsilon',K,\mathsf{d}')$$

so $h_{\mathsf{d}}(T,K)\leqslant h_{\mathsf{d}'}(T,K)$, which proves the lemma by symmetry. □

Lemma 3.3. *Let T be a uniformly continuous map on a locally compact σ-compact metric space (X,d). For compact sets $K_1,K_2\subseteq X$ and $K\subseteq K_1\cup K_2$,*

$$h_{\mathsf{d}}(T,K)\leqslant\max\{h_{\mathsf{d}}(T,K_1),h_{\mathsf{d}}(T,K_2)\}.$$

Proof. Clearly $s_n(\varepsilon,K,\mathsf{d})\leqslant s_n(\varepsilon,K_1,\mathsf{d})+s_n(\varepsilon,K_2,\mathsf{d})$ for any n and $\varepsilon>0$. Fix $\varepsilon>0$, and choose $i(n)\in\{1,2\}$ to have

$$\max\{s_n(\varepsilon,K_1,\mathsf{d}),s_n(\varepsilon,K_2,\mathsf{d})\}=s_n(\varepsilon,K_{i(n)},\mathsf{d})$$

for each $n \geqslant 1$, so that $s_n(\varepsilon, K, \mathsf{d}) \leqslant 2s_n(\varepsilon, K_{i(n)}, \mathsf{d})$. By choosing a subsequence along which there is convergence to the limit superior, and then choosing a further subsequence along which the function $n \mapsto i(n)$ is constant, we find $n_j \to \infty$ with

$$\frac{1}{n_j} \log s_{n_j}(\varepsilon, K, \mathsf{d}) \to \limsup_{n\to\infty} \frac{1}{n} \log s_n(\varepsilon, K, \mathsf{d})$$

as $j \to \infty$, and with $i(n_j) = i^* \in \{1, 2\}$ for all $j \geqslant 1$. Then

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log s_n(\varepsilon, K, \mathsf{d}) &= \lim_{j\to\infty} \frac{1}{n_j} \log s_{n_j}(\varepsilon, K, \mathsf{d}) \\
&\leqslant \lim_{j\to\infty} \frac{1}{n_j} \left(\log 2 + \log s_{n_j}(\varepsilon, K_{i^*}, \mathsf{d})\right) \\
&\leqslant \limsup_{n\to\infty} \frac{1}{n} \log s_n(\varepsilon, K_{i^*}, \mathsf{d}) \\
&\leqslant h_{\mathsf{d}}(T, K_{i^*}). \qquad (3.3)
\end{aligned}$$

Now choose a sequence $\varepsilon_k \to 0$ as $k \to \infty$ with i^* (which depends on ε) constant for all $k \geqslant 1$. Then the inequality (3.3) shows that

$$h_{\mathsf{d}}(T, K) \leqslant h_{\mathsf{d}}(T, K_{i^*}) \leqslant \max\{h_{\mathsf{d}}(T, K_1), h_{\mathsf{d}}(T, K_2)\}.$$

□

Corollary 3.4. *For any $\delta > 0$, the supremum of $h_{\mathsf{d}}(T, K)$ over all compact sets K of diameter no more than δ coincides with $h_{\mathsf{d}}(T)$.*

Proof. Any compact set $K' \subseteq X$ has a finite cover

$$K' \subseteq B_{\delta/2}(x_1) \cup \cdots \cup B_{\delta/2}(x_k)$$

by metric open balls of radius $\delta/2$. By Lemma 3.3 and induction,

$$h_{\mathsf{d}}(T, K') \leqslant \max_{1\leqslant j\leqslant k} \{h_{\mathsf{d}}(T, K' \cap \overline{B_{\delta/2}(x_j)})\},$$

completing the proof. □

If $T : (X, \mathsf{d}) \to (X, \mathsf{d})$ is a continuous map on a compact metric space and $\varepsilon > 0$ is given, then equations (2.2) and (2.3) show that if $\mathscr{U}$ is the cover of X by all open balls of radius 2ε, and $\mathscr{V}$ is any cover by open balls of radius $\varepsilon/2$, then

$$N\left(\bigvee_{j=0}^{n-1} T^{-j}\mathscr{U}\right) \leqslant r_n^T(\varepsilon, X, \mathsf{d}) \leqslant s_n^T(\varepsilon, X, \mathsf{d}) \leqslant N\left(\bigvee_{j=0}^{n-1} T^{-j}\mathscr{V}\right). \qquad (3.4)$$

This gives several useful results. As $\varepsilon \to 0$ in equation (3.4), the outer terms converge to $h_{\text{cover}}(T)$ by Proposition 2.15, so

$$h_{\text{top}}(T) = h_{\mathsf{d}}(T).$$

Moreover, Proposition 2.15 and the inequality (3.4) together show that

$$h_{\mathsf{d}}(T) = \lim_{\varepsilon\to 0} \liminf_{n\to\infty} \frac{1}{n} \log r_n^T(\varepsilon, X, \mathsf{d}) = \lim_{\varepsilon\to 0} \liminf_{n\to\infty} \frac{1}{n} \log s_n^T(\varepsilon, X, \mathsf{d}). \quad (3.5)$$

As mentioned in Lemmas 2.19, 2.20 and in Exercise 2.2.3, topological entropy has functorial properties[19]. A small complication does arise in the non-compact setting, which will be explained in Lemma 3.7. In the proofs we will find spanning and separating sets for various maps, so we add a superscript to r and s to denote this when needed.

Lemma 3.5. *Let $T : X \to X$ be a uniformly continuous map on a metric space. Then $h_{\mathsf{d}}(T^k) = kh_{\mathsf{d}}(T)$ for any $k \geqslant 1$.*

Proof. Let K be a compact subset of X, and write d for the metric on X. Clearly $r_n^{T^k}(\varepsilon, K, \mathsf{d}) \leqslant r_{kn}^T(\varepsilon, K, \mathsf{d})$ so $h_{\mathsf{d}}(T^k) \leqslant kh_{\mathsf{d}}(T)$. For the reverse inequality, given any $\varepsilon > 0$ we may choose (by uniform continuity) a $\delta > 0$ such that

$$\mathsf{d}(x, y) \leqslant \delta \implies \mathsf{d}(T^j x, T^j y) < \varepsilon, \quad 0 \leqslant j < k.$$

It follows that an (n, δ)-spanning set for K for the map T^k is automatically a (kn, ε)-spanning set for the map T, and so $r_{kn}^T(\varepsilon, K, \mathsf{d}) \leqslant r_n^{T^k}(\delta, K)$, and hence $h_{\mathsf{d}}(T^k) \geqslant kh_{\mathsf{d}}(T)$. □

Corollary 3.6. *Let $T : X \to X$ be a continuous map on a compact metric space. Then $h_{\text{top}}(T^k) = kh_{\text{top}}(T)$ for any $k \geqslant 1$.*

Lemma 3.7. *If $T_i : X_i \to X_i$ are uniformly continuous maps of metric spaces (X_i, d_i) for $i = 1, 2$, then*

$$h_{\mathsf{d}}(T_1 \times T_2) \leqslant h_{\mathsf{d}_1}(T_1) + h_{\mathsf{d}_2}(T_2),$$

where the metric on $X_1 \times X_2$ is

$$\mathsf{d}\left((x_1, x_2), (y_1, y_2)\right) = \max\{\mathsf{d}_1(x_1, y_1), \mathsf{d}_2(x_2, y_2)\}.$$

If the limit supremums in Definition 3.1 *are limits or if X_1 is compact, then*

$$h_{\mathsf{d}}(T_1 \times T_2) = h_{\mathsf{d}_1}(T_1) + h_{\mathsf{d}_2}(T_2).$$

Proof. Let K_i be compact in X_i, and let E_i be an (n, ε)-spanning set for K_i. Then $E_1 \times E_2$ is (n, ε)-spanning for $K_1 \times K_2$ with respect to $T_1 \times T_2$. It follows that

$$r_n^{T_1\times T_2}(\varepsilon, K_1\times K_2, \mathsf{d}) \leqslant r_n^{T_1}(\varepsilon, K_1, \mathsf{d}_1) r_n^{T_2}(\varepsilon, K_2, \mathsf{d}_2),$$

so

$$h_{\mathsf{d}}(T_1\times T_2, K_1\times K_2) \leqslant h_{\mathsf{d}_1}(T_1, K_1) + h_{\mathsf{d}_2}(T_2, K_2).$$

For the reverse inequality, we assume that we have convergence as $n\to\infty$ in Definition 3.1. If F_i is (n,ε)-separated for K_i under T_i, then $F_1\times F_2$ is (n,ε)-separated for $K_1\times K_2$ under $T_1\times T_2$, so

$$s_n^{T_1\times T_2}(\varepsilon, K_1\times K_2, \mathsf{d}) \geqslant s_n^{T_1}(\varepsilon, K_1, \mathsf{d}_1) s_n^{T_2}(\varepsilon, K_2, \mathsf{d}_2). \tag{3.6}$$

By the assumption that the limit in n exists, we deduce that

$$h_{\mathsf{d}}(T_1\times T_2, K_1\times K_2) \geqslant h_{\mathsf{d}_1}(T_1, K_1) + h_{\mathsf{d}_2}(T_2, K_2).$$

Now write $\pi_i : X_1\times X_2 \to X_i$ for the projection map. Then for any compact set $K\subseteq X_1\times X_2$, $\pi_1(K)$ and $\pi_2(K)$ are compact and $K\subseteq \pi_1(K)\times\pi_2(K)$, so

$$\begin{aligned} h_{\mathsf{d}}(T_1\times T_2) &= \sup_{K\subseteq X_1\times X_2} h_{\mathsf{d}}(T_1\times T_2, K) \\ &= \sup_{K_i\subseteq X_i} h_{\mathsf{d}}(T_1\times T_2, K_1\times K_2) \\ &= \sup_{K_1\subseteq X_1} h_{\mathsf{d}_1}(T_1, K_1) + \sup_{K_2\subseteq X_2} h_{\mathsf{d}_2}(T_2, K_2) = h_{\mathsf{d}_1}(T_1) + h_{\mathsf{d}_2}(T_2) \end{aligned}$$

by Lemma 3.3.

Finally, suppose that X_1 is compact. Any compact set in $X_1\times X_2$ is a subset of $X_1\times K_2$ for some compact set $K_2\subseteq X_2$, so

$$h_{\mathsf{d}}(T_1\times T_2) = \sup_{K_2\subseteq X_2} \{h_{\mathsf{d}}(T_1\times T_2, X_1\times K_2).$$

Fix a compact set $K_2\subseteq X_2$ and some $\delta > 0$. By equation (3.6) applied with $K_1 = X_1$, we have

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n}\log s_n^{T_1\times T_2}(\varepsilon, K_1\times K_2, \mathsf{d}) &\geqslant \limsup_{n\to\infty} \frac{1}{n}\left(\log s_n^{T_1}(\varepsilon, K_1, \mathsf{d}_1)\right. \\ &\qquad \left. + \log s_n^{T_2}(\varepsilon, K_2, \mathsf{d}_2)\right) \\ &\geqslant \liminf_{n\to\infty} \frac{1}{n}\log s_n^{T_1}(\varepsilon, K_1, \mathsf{d}_1) \\ &\qquad + \limsup_{n\to\infty} \frac{1}{n}\log s_n^{T_2}(\varepsilon, K_2, \mathsf{d}_2). \end{aligned}$$

Thus $h_{\mathsf{d}}(T_1\times T_2, X_1\times K_2) \geqslant h_{\mathsf{d}_1}(T_1) + h_d(T_2, K_2)$ by equation (3.5), completing the proof. □

One of the ways in which topological entropy for maps on non-compact spaces will be useful is to linearize certain entropy calculations, initially

on compact spaces. A simple example, which will be used in proving Theorem 3.12, involves lifting a toral automorphism: The r-torus $\mathbb{T}^r$ may be thought of as the quotient group $\mathbb{R}^r/\mathbb{Z}^r$, and the usual Euclidean metric d on $\mathbb{R}^r$ induces a metric d' on $\mathbb{T}^r$ by setting

$$\mathsf{d}'(x+\mathbb{Z}^r, y+\mathbb{Z}^r) = \min_{n\in\mathbb{Z}^r} \mathsf{d}(x, y+n).$$

Notice that this makes the quotient map $\pi : \mathbb{R}^r \to \mathbb{T}^r$ into a local isometry: every point in $\mathbb{R}^r$ has a neighborhood that is mapped isometrically onto an open set in $\mathbb{T}^r$. The next result gives general conditions under which entropy is preserved in this kind of lift.

Proposition 3.8. *Let $\pi : (X, \mathsf{d}) \to (X', \mathsf{d}')$ be a continuous surjective map with the property that for some $\delta > 0$ the map π restricted to $B_{\delta,\mathsf{d}}(x)$ is an isometric surjection onto $B_{\delta,\mathsf{d}'}(\pi(x))$ for every $x \in X$. If*

$$\begin{array}{ccc} (X,\mathsf{d}) & \xrightarrow{\ T\ } & (X,\mathsf{d}) \\ {\scriptstyle\pi}\big\downarrow & & \big\downarrow{\scriptstyle\pi} \\ (X',\mathsf{d}') & \xrightarrow[\ T'\ ]{} & (X',\mathsf{d}') \end{array}$$

is a commutative diagram and both T and T' are uniformly continuous, then $h_{\mathsf{d}}(T) = h_{\mathsf{d}'}(T')$.

Proof. If $K \subseteq X$ is compact with $\operatorname{diam}(K) < \delta$ then by the local isometry property the image $\pi(K)$ is compact with $\operatorname{diam}(K) < \delta$; moreover any compact set $K' \subseteq X'$ with $\operatorname{diam}(K') < \delta$ is an image of such a set. Choose (by uniform continuity) an $\varepsilon \in (0, \delta)$ so that $\mathsf{d}(x, y) \leqslant \varepsilon$ implies $\mathsf{d}(Tx, Ty) < \delta$, and let $E \subseteq K$ be an (n, ε) separating set for T. We claim that $\pi(E) \subseteq \pi(K)$ is an (n, ε)-separated set for T' of equal cardinality. By the local isometry property, if $x, y \in E$ are distinct then $\pi(x) \neq \pi(y)$; given such a pair choose if possible j so that $\max_{i\leqslant j} \mathsf{d}(T^i x, T^i y) \leqslant \varepsilon$ but $\mathsf{d}(T^{j+1}x, T^{j+1}y) > \varepsilon$. By the choice of ε, $\mathsf{d}(T^{j+1}x, T^{j+1}y) < \delta$ so by the local isometry property $\mathsf{d}'(T'^{j+1}\pi(x), T'^{j+1}\pi(y)) = \mathsf{d}(T^{j+1}x, T^{j+1}y) > \varepsilon$. If $\mathsf{d}(x, y) > \varepsilon$, then there is no such j; in this case use the fact that $\operatorname{diam}(E) < \delta$ to see that $\pi(x)$ and $\pi(y)$ are again ε-separated in $\pi(E)$. Thus $\pi(E)$ is (n, ε)-separated for T'. It follows that $s_n(\varepsilon, K, \mathsf{d}) \leqslant s_n(\varepsilon, \pi(K), \mathsf{d}')$. Conversely, if E' is an (n, ε)-separated under T' for $\pi(K)$, where $K \subseteq X$ is compact with $\operatorname{diam}(K) < \delta$ then for distinct $x, y \in \pi^{-1}E' \cap K$, $\mathsf{d}'(T'^i\pi(x), T'^i\pi(y)) > \varepsilon$ implies $d(T^i, T^i y) > \varepsilon$, so $\pi^{-1}E' \cap K$ is (n, ε)-separated for T. Thus $s_n(\varepsilon, \pi(K), T', \mathsf{d}') \leqslant s_n(\varepsilon, K, T, \mathsf{d})$, so they are equal. This shows that $h_{\mathsf{d}}(T, K) = h_{\mathsf{d}'}(T', \pi(K))$ for all small enough compact sets, and Corollary 3.4 shows that $h_{\mathsf{d}}(T) = h_{\mathsf{d}'}(T')$. □

Exercises for Section 3.1

Exercise 3.1.1. Show that if (X, d) is compact, then equivalent metrics are uniformly equivalent and continuous maps are uniformly continuous. Deduce that in this case $h_{\mathsf{d}}(T) = h_{\text{top}}(T)$.

Exercise 3.1.2. For the map $x \mapsto 2x \pmod 1$ compute $h_{\text{top}}(T)$ directly from the definition using separating sets in $\mathbb{T}$, and then compute it using Proposition 3.8.

Exercise 3.1.3. Assume that the Proposition 3.8 is weakened slightly to say that for each $x \in X$ there is a $\delta = \delta(x) > 0$ with the property that the map π restricted to $B_{\delta,\mathsf{d}}(x)$ is an isometric surjection onto $B_{\delta,\mathsf{d}'}(\pi(x))$. Show that in this case we can deduce that $h_{\mathsf{d}}(T) \geqslant h_{\mathsf{d}'}(T')$.

Exercise 3.1.4. Show that the canonical projection map

$$\pi : \mathrm{SL}_2(\mathbb{R}) \to X = \mathrm{SL}_2(\mathbb{Z}) \backslash \mathrm{SL}_2(\mathbb{R})$$

satisfies the weaker hypothesis in Exercise 3.1.3 but does not satisfy the hypothesis of Proposition 3.8. In this and similar cases, equality of topological entropy does nonetheless hold for certain algebraic maps.
(a) Show that the horocycle time-1 map, defined on X by

$$\mathrm{SL}_2(\mathbb{Z})g \longmapsto \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \mathrm{SL}_2(\mathbb{Z})g$$

has zero entropy.
(b) Compute the entropy of the time-1 geodesic flow, defined on X by

$$\mathrm{SL}_2(\mathbb{Z})g \longmapsto \begin{pmatrix} \mathrm{e}^{1/2} & 0 \\ 0 & \mathrm{e}^{-1/2} \end{pmatrix} \mathrm{SL}_2(\mathbb{Z})g.$$

3.2 Homogeneous Measures

Recall that $T : (X, \mathsf{d}) \to (X, \mathsf{d})$ is a uniformly continuous map on a locally compact metric space; since the metric d is fixed we suppress it in the notation.

Definition 3.9. *A* Bowen ball about x *is a set of the form*

$$D_n(x, \varepsilon, T) = \bigcap_{k=0}^{n-1} T^{-k} \left(B_\varepsilon(T^k x) \right),$$

where $B_\varepsilon(y) = \{z \in X \mid \mathsf{d}(y, z) < \varepsilon\}$ *is the metric open ball around* y *of radius* ε. *A Borel measure* μ *on* X *is called* T-homogeneous *if*

(1) $\mu(K) < \infty$ *for all compact sets* $K \subseteq X$,

(2) $\mu(K) > 0$ *for some compact set* $K \subseteq X$,
(3) *for each* $\varepsilon > 0$ *there is a* $\delta > 0$ *and a* $c > 0$ *with the property that*

$$\mu\left(D_n(y, \delta, T)\right) \leqslant c\mu\left(D_n(x, \varepsilon, T)\right)$$

for all $n \geqslant 0$ *and* $x, y \in X$.

We shall see later that T-homogeneous measures exist in many useful situations; on the other hand they are so special that the rate of the decay of the measure (with respect to any T-homogeneous measure) of the Bowen balls computes the topological entropy. Write

$$k_{\mathsf{d}}(\mu, T) = \lim_{\varepsilon \searrow 0} \limsup_{n \to \infty} -\frac{1}{n} \log \mu\left(D_n(y, \varepsilon, T)\right); \tag{3.7}$$

notice that property (3) of Definition 3.9 implies that the quantity in equation (3.7) is independent of y.

Theorem 3.10 (Bowen). *Let* $T : (X, \mathsf{d}) \to (X, \mathsf{d})$ *be a uniformly continuous map on a locally compact metric space and let* μ *be a* T*-homogeneous Borel measure on* X. *Then*

$$h_{\mathsf{d}}(T) = k_{\mathsf{d}}(\mu, T).$$

If X *is compact,* $\mu(X) = 1$, *and* μ *is* T*-invariant and* T*-homogeneous, then*

$$h_\mu(T) = k_{\mathsf{d}}(\mu, T).$$

Proof. Let $K \subseteq X$ be compact. By property (2) of Definition 3.9 and the local compactness of X, there is an open set $U \supseteq K$ with $\mu(U) < \infty$; choose $\varepsilon > 0$ small enough to ensure that

$$B_\varepsilon(K) = \{x \in X \mid \inf_{y \in K} \mathsf{d}(x, y) < \varepsilon\} \subseteq U.$$

Let $E \subseteq K$ be an (n, ε)-separated set of maximal cardinality. For distinct points $x_1, x_2 \in E$ the sets $D_n(x_1, \varepsilon/2, T)$ and $D_n(x_2, \varepsilon/2, T)$ are disjoint and

$$\bigsqcup_{x \in E} D_n(x, \varepsilon/2, T) \subseteq U$$

is a disjoint union. By property (3) there are constants δ, c with

$$\mu\left(D_n(y, \delta, T)\right) \leqslant c\mu\left(D_n(x, \varepsilon/2, T)\right) \text{ for all } x, y.$$

Thus for a fixed y we have

$$\begin{aligned}\mu\left(D_n(y, \delta, T)\right) s_n(\varepsilon, K) &\leqslant \sum_{x \in E} c\mu\left(D_n(x, \varepsilon/2, T)\right) \\ &= c\mu\left(\bigsqcup_{x \in E} D_n(x, \varepsilon/2, T)\right) \leqslant c\mu(U),\end{aligned}$$

so

$$\limsup_{n\to\infty} \frac{1}{n} \log s_n(\varepsilon, K, \mathsf{d}) \leqslant \limsup_{n\to\infty} -\frac{1}{n} \log \mu\left(D_n(y, \delta, T, \mathsf{d})\right).$$

Taking $\varepsilon \to 0$ gives $h_{\mathsf{d}}(T, K) \leqslant k_{\mathsf{d}}(\mu, T)$, so $h_{\mathsf{d}}(T) \leqslant k_{\mathsf{d}}(\mu, T)$.

The reverse inequality is similar: let K be a given compact set and let $\varepsilon > 0$ be given. We may assume that $\mu(K) > 0$. Choose $\delta > 0$ and $c > 0$ as in property (3), and let F be an (n, δ)-spanning set for K. Then, by definition of a spanning set,

$$\bigcup_{x\in F} D_n(x, \delta, T) \supseteq K,$$

which together with property (3) implies that

$$c\mu\left(D_n(y, \varepsilon, T)\right) r_n(\delta, K, \mathsf{d}) \geqslant \mu(K) > 0.$$

Therefore

$$\limsup_{n\to\infty} \frac{1}{n} \log r_n(\delta, K, \mathsf{d}) \geqslant \limsup -\frac{1}{n} \log \mu\left(D_n(y, \varepsilon, T)\right)$$

Taking $\varepsilon \to 0$ gives $h_{\mathsf{d}}(T, K) \geqslant k_{\mathsf{d}}(\mu, T)$, so $h_{\mathsf{d}}(T) = k_{\mathsf{d}}(\mu, T)$.

Now assume that X is compact and μ is a T-invariant, T-homogeneous probability measure. By the variational principle (Theorem 2.22),

$$h_\mu(T) \leqslant h_{\mathrm{top}}(T) = h_{\mathsf{d}}(T) = k_{\mathsf{d}}(\mu, T).$$

For the reverse inequality, fix $\varepsilon > 0$ and choose $\delta > 0$, $c > 0$ with the property that

$$\mu\left(D_n(x, \delta, T)\right) \leqslant c\mu\left(D_n(y, \varepsilon, T)\right)$$

for all $x, y \in X$ and all $n \geqslant 0$, and let $\xi = \{A_1, \ldots, A_r\}$ be a measurable partition of X into sets of diameter no more than δ. Then for x in an atom A of $\bigvee_{k=0}^{n-1} T^{-k}\xi = \xi_n$, we have $A \subseteq D_n(x, \delta, T)$, and so $\mu(A) \leqslant c\mu\left(D_n(y, \varepsilon, T)\right)$. It follows that

$$\begin{aligned} H\left(\bigvee_{k=0}^{n-1} T^{-k}\xi\right) &= -\sum_{A\in\xi_n} \mu(A) \log \mu(A) \\ &\geqslant -\sum_{A\in\xi_n} \mu(A) \log c\mu\left(D_n(y, \varepsilon, T)\right) \\ &= -\log c - \log \mu\left(D_n(y, \varepsilon, T)\right). \end{aligned}$$

Thus

$$h_\mu(T) \geqslant h_\mu(T, \xi) \geqslant \limsup_{n\to\infty} -\frac{1}{n} \log \mu\left(D_n(y, \varepsilon, T)\right);$$

taking $\varepsilon \to \infty$ shows that $h_\mu(T) \geqslant k_{\mathsf{d}}(\mu, T)$. □

We have seen that topological entropy for continuous maps on compact metric spaces adds over products in Lemma 2.20. In general the topological

entropy for uniformly continuous maps on a non-compact space does not add over products without an additional assumption (as seen in Lemma 3.7, where we give one such instance). Here we record two such situations in which we recover additivity, in the context of homogeneous measures.

Lemma 3.11. *Let $T_i : (X_i, \mathsf{d}_i) \to (X_i, \mathsf{d}_i)$ be uniformly continuous with homogeneous measure μ_i for $i = 1, 2$, and let $\mathsf{d} = \max\{\mathsf{d}_1, \mathsf{d}_2\}$.*

(1) *If $-\frac{1}{n} \log \mu_1 \left(D_n(y, \varepsilon, T_1)\right)$ converges for any $y \in X_1$ then*

$$h_{\mathsf{d}}(T_1 \times T_2) = h_{\mathsf{d}_1}(T_1) + h_{\mathsf{d}_2}(T_2).$$

(2) *For $T_1 \times T_1$ on $X_1 \times X_1$, we have $h_{\mathsf{d}}(T_1 \times T_1) = 2h_{\mathsf{d}_1}(T_1)$.*

Proof. If μ_i is T_i-homogeneous, then $\mu_1 \times \mu_2$ is $T_1 \times T_2$-homogeneous with respect to the maximum metric d on $X_1 \times X_2$. Thus (1) follows from Lemma 3.7, and (2) follows since we may simply restrict attention to that sequence of ns realizing the limit supremum in both systems simultaneously, and then we may assume convergence. □

Exercises for Section 3.2

Exercise 3.2.1. (a) For the time-1 map T of the geodesic flow on defined as in Exercise 3.1.4 and lifted to $\mathrm{SL}_2(\mathbb{R})$, compute $k_{\mathsf{d}}(\mu, T)$ and $h_{\mathsf{d}}(T)$.
(b) Compute $k_{\mathsf{d}}(\mu, T)$ for the time-1 map of the horocycle flow, and deduce that the horocycle flow has zero entropy.
(c) Use Proposition 3.8 to find the entropy of the horocycle flow on $\Gamma \backslash \mathrm{SL}_2(\mathbb{R})$ for any uniform lattice Γ (that is, with compact quotient).

3.3 Calculating Topological Entropy

For continuous maps that are highly homogeneous (that is, their action on each part of the space looks the same) Section 3.2 shows that it is possible to compute[20] the topological entropy locally.

3.3.1 Toral Automorphisms

We are now in a position to compute the topological entropy of a linear map on $\mathbb{R}^n$, and this will allow the entropy of a toral automorphism to be deduced. In addition, the connection between the decay in volume of the Bowen balls and the topological entropy will at the same time show that Lebesgue measure on the torus is maximal for toral automorphisms. We saw in Section 4.5 a specific example of a toral automorphism with the property that Lebesgue

measure is the unique measure of maximal entropy. This is true in general, but the methods developed in this section do not show that. More robust ways to understand entropy will be developed later, and unique maximality will be shown there.

Theorem 3.12. *Let $A : \mathbb{R}^r \to \mathbb{R}^r$ be the linear automorphism defined by the matrix $A \in \mathrm{GL}_r(\mathbb{R})$. Then, if d is the usual Euclidean metric on $\mathbb{R}^r$,*

$$h_{\mathsf{d}}(A) = \sum_{\lambda} \log^+ |\lambda| \tag{3.8}$$

where the sum is taken (with multiplicities) over all the eigenvalues of A, and $\log^+(x) = \max\{\log x, 0\}$.

Before proving this, we note a particularly simple instance. If A is a real diagonal matrix,

$$A = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_r \end{pmatrix},$$

then we may replace d with the uniformly equivalent metric

$$\mathsf{d}'(x, y) = \max_{1 \leqslant i \leqslant r} \{|x_i - y_i|\},$$

where $x = (x_1, \ldots, x_r)^t$, so that $B_{\mathsf{d}',\varepsilon}(0)$ is an r-dimensional cube with side 2ε centered at 0. Since A is diagonal, the map $x \mapsto A^{-1}x$ dilates the ith axis by the factor λ_i^{-1}; this is illustrated in Figure 3.1 for the situation

$$r = 3, 0 < \lambda_1 < 1, \lambda_2 > 1, \lambda_3 > 1.$$

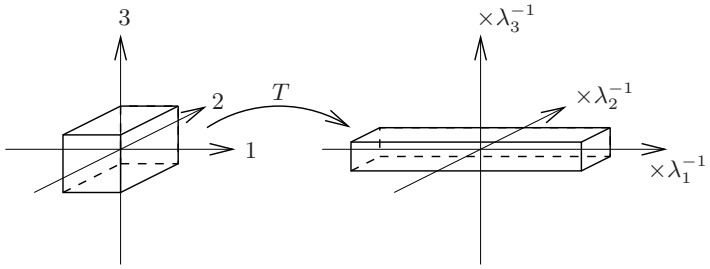

Fig. 3.1. Action of A^{-1} with $0 < \lambda_1 < 1$, $\lambda_2 > 1$, $\lambda_3 > 1$.

Thus the Bowen ball $D_n(0, \varepsilon, A, \mathsf{d}')$ is an r-dimensional rectangular parallelepiped; the ith side has length 2ε if $|\lambda_i| \leqslant 1$ and length $2\varepsilon|\lambda_i^{-n+1}|$ if $|\lambda_i| > 1$. It follows that

$$m\left(D_n(0,\varepsilon,A,\mathsf{d}')\right) = (2\varepsilon)^r \prod_{|\lambda_i|>1} \lambda_i^{-n+1},$$

which shows that in this case

$$k_{\mathsf{d}'}(m,A) = \sum_{\lambda} \log^+ |\lambda|,$$

giving the claimed formula by Theorem 3.10.

This simple case should be born in mind in the general case below. There are two difficulties to overcome. First, the matrix A may have complex eigenvalues. Second, these eigenvalues may give rise to non-trivial Jordan blocks (that is, the matrix might not be diagonalizable). The first problem is easily dealt with by using a complex vector space, identifying[21] for example the action of $\begin{pmatrix} a & b \\ -b & a \end{pmatrix}$ (or any of its conjugates) on $\mathbb{R}^2$ with multiplication by $a+\mathrm{i}b$ on $\mathbb{C}$. The second involves an important principle which will arise several times: Jordan blocks distort the exponentially decaying parallelepiped by an amount that is polynomially bounded. In the limit it is the exponential rate that determines the volume decay, and that is the essence of the formula in Theorem 3.12.

Proof of Theorem 3.12. First notice that $x \mapsto Ax$ is uniformly continuous with respect to d, and Lebesgue measure m on $\mathbb{R}^r$ is A-homogeneous, since both the metric and the measure are translation invariant. Thus by Theorem 3.10

$$h_{\mathsf{d}}(A) = k_{\mathsf{d}}(m,A).$$

By choosing a suitable basis in $\mathbb{R}^r$ we may assume that the matrix A has the Jordan form

$$A = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_s \end{pmatrix}$$

where each block J_i corresponds to an eigenvalue λ_i of A. In the case $\lambda_i \in \mathbb{R}$, the corresponding block has the form

$$J = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & 1 & \\ & & \ddots & \ddots \\ & & & \lambda_i \ 1 \end{pmatrix} \tag{3.9}$$

and in the case of a pair of eigenvalues $\lambda_i, \overline{\lambda_i} \in \mathbb{C}$, the corresponding block has the form

$$J = \begin{pmatrix} \Lambda_i & I_2 & & \\ & \Lambda_i & I_2 & \\ & & \ddots & \ddots \\ & & & \Lambda_i \ I_2 \end{pmatrix}, \tag{3.10}$$

where $\lambda_i = a + ib$, $\Lambda_i = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}$, and $I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

Since the Lebesgue measure in $\mathbb{R}^r$ is the product of the Lebesgue measures on the subspace corresponding to each block, it is sufficient to prove equation (3.8) on each block separately, so long as we also establish that the limit

$$\lim_{n\to\infty} -\frac{1}{n} \log m_i \left(D_n(0, \varepsilon, J_i, \mathsf{d}_i)\right)$$

exists, so that the entropy adds over the product by Lemma 3.11(1), where d_i denotes the Euclidean metric on the ith block.

Thus we are reduced to a single Jordan block $J = J_i$. If it corresponds to a complex conjugate pair of eigenvalues as in equation (3.10), then by making the identification $(x, y) \mapsto x + \mathrm{i}y$ between $\mathbb{R}^2$ and $\mathbb{C}$ we may assume that the block always has the form in equation (3.9), with $\lambda_i \in \mathbb{R}$ in the real case and $\lambda_i \in \mathbb{C}$ in the complex case. So assume that

$$J = \begin{pmatrix} \lambda & 1 & & \\ & \lambda & 1 & \\ & & \ddots & \ddots \\ & & & \lambda & 1 \end{pmatrix} = \lambda I_\ell + N$$

is an $\ell \times \ell$ matrix corresponding to one Jordan block, acting on $\mathbb{K}^\ell$ where $\mathbb{K}$ is either $\mathbb{R}$ or $\mathbb{C}$ as appropriate. It is easily checked by induction that

$$J^n = (\lambda I + N) = \sum_{k=0}^{\ell-1} \binom{n}{k} \lambda^{n-k} N^k,$$

for any $n \in \mathbb{Z}$, where $\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!}$ for all $n \in \mathbb{Z}$ and $k \geqslant 0$, and

$$(N^m)_{i,j} = \begin{cases} 1 & \text{if } (i,j) = (1, m+1), (2, m+2), \ldots, (\ell - m, \ell); \\ 0 & \text{if not;} \end{cases}$$

and N^m is the zero matrix if $m \geqslant \ell$.

Fix $\varepsilon > 0$ and $\delta > 0$, and define $t = (1 + \delta)|\lambda|$. Then every entry of the matrix $t^{-n} J^n$ has the form $\frac{|\lambda|^n}{t^n}$ multiplied by a polynomial in n, so (thinking of the matrix $t^{-n} J^n$ as an operator on $\mathbb{K}^\ell$ with respect to any norm) there is some constant $C > 0$ such that

$$\|t^{-n} J^n\|_{\text{operator}} \leqslant C$$

for all $n \geqslant 0$. It follows that if d is a metric on $\mathbb{K}^\ell$ and $\mathsf{d}(t^i x, 0) < \frac{\varepsilon}{C}$ for some $i \geqslant 0$, then $\mathsf{d}(J^i x, 0) < \varepsilon$. Thus

$$D_n\left(0, \tfrac{\varepsilon}{C}, t, \mathsf{d}\right) \subseteq D_n\left(0, \varepsilon, J, \mathsf{d}\right),$$

and

$$D_n\left(0, \tfrac{\varepsilon}{C}, t, \mathsf{d}\right) = \begin{cases} t^{-n+1} B_{\varepsilon/C}(0) & \text{if } t \geqslant 1, \\ B_{\varepsilon/C}(0) & \text{if } t < 1. \end{cases}$$

It follows that

$$[\mathbb{K} : \mathbb{R}]\ell \log^+ (|\lambda|(1+\delta)) = [\mathbb{K} : \mathbb{R}]\ell \log^+ t \geqslant \limsup_{n\to\infty} -\frac{1}{n} \log m \left(D_n(0, \varepsilon, J, \mathsf{d})\right) \tag{3.11}$$

where m is Lebesgue measure on $\mathbb{K}^\ell$.

Now let $t_1 = \frac{|\lambda|}{1+\delta} < |\lambda|$, so that the entries of the matrix $t_1^n J^{-n}$ are of the form $\frac{t_1^n}{|\lambda|^n} = \frac{1}{(1+\delta)^n}$ multiplied by a polynomial in n for all $n \geqslant 1$, so that we have some $C > 0$ with

$$\|t_1^n J^{-n}\|_{\text{operator}} \leqslant C$$

for all $n \geqslant 0$. As before, it follows that

$$D_n\left(0, \varepsilon, J, \mathsf{d}\right) \subseteq D_n\left(0, C\varepsilon, t_1, \mathsf{d}\right)$$

and therefore

$$\liminf_{n\to\infty} -\frac{1}{n} \log m \left(D_n(0, \varepsilon, J, \mathsf{d})\right) \geqslant [\mathbb{K} : \mathbb{R}]\ell \log^+ t_1 = [\mathbb{K} : \mathbb{R}]\ell \log^+ \frac{|\lambda|}{1+\delta}.$$

Together with equation (3.11), this shows (by taking $\delta \to 0$) that

$$\lim_{n\to\infty} -\frac{1}{n} \log m \left(D_n(0, \varepsilon, J, \mathsf{d})\right) = [\mathbb{K} : \mathbb{R}]\ell \log^+ |\lambda|,$$

concluding the proof of Theorem 3.12. □

Theorem 3.13. *The entropy of the automorphism T_A of the r-torus $\mathbb{T}^r$ associated to a matrix $A \in \mathrm{GL}_r(\mathbb{Z})$ is given by*

$$h_{\text{top}}(T_A) = h_m(T_A) = \sum_\lambda \log^+ |\lambda|$$

where m denotes Lebesgue measure, the sum is taken (with multiplicities) over all the eigenvalues of A, and $\log^+(x) = \max\{\log x, 0\}$.

Proof. By Theorem 2.18, $h_{\text{top}}(T_A) = h_{\mathsf{d}}(T_A)$ where d is the metric induced on $\mathbb{T}^r$ by the usual metric on $\mathbb{R}^r$. By Proposition 3.8, it is enough to compute the topological entropy of the map lifted to $\mathbb{R}^r$, and Theorem 3.12 gives the formula. Finally Theorem 3.10 shows that the measure-theoretic entropy with respect to Lebesgue measure has the same value, since Lebesgue measure is T_A-homogeneous (as both the Lebesgue measure and the metric are translation invariant). □

3.3.2 Automorphisms of Solenoids

With the machinery developed in Section 3.3 and the adelic language from Appendix B it is relatively straightforward to extend Theorem 3.13 to automorphisms of solenoids[22], which is the main step in computing the entropy of any group automorphism.

Definition 3.14. *A* solenoid *is a finite-dimensional, connected, compact, abelian group. Equivalently, its dual group is a finite rank, torsion-free, discrete abelian group, and thus is a subgroup of $\mathbb{Q}^d$ for some $d \geqslant 1$.*

A formula for the topological entropy of an automorphism of a solenoid was found by Yuzvinskiĭ [148]; in this section we use adeles to give a simple proof of Yuzvinskiĭ's formula[23]. In particular, this approach shows that the topological entropy of an automorphism of a solenoid is made up of geometrical contributions in exactly the same way as is the case for toral automorphisms.

Let T be an automorphism of an r-dimensional solenoid X, so the dual automorphism $\widehat{T} : \widehat{X} \to \widehat{X}$ extends to an automorphism of $\mathbb{Q}^r$, which may be described by an element of $\mathrm{GL}_r(\mathbb{Q})$, which we denote by A and write $T = T_A$. In the toral case $X = \mathbb{T}^r$, the matrix A lies in $\mathrm{GL}_r(\mathbb{Z})$, and as shown in Theorem 3.13 we have

$$h_{\mathrm{top}}(T_A) = \sum_{|\lambda|>1} \log|\lambda|,$$

where the sum is taken over the set of eigenvalues of A, with multiplicities. Moreover, as seen in the proof of Theorem 3.13 (and, in a different setting, in Section 4.5) an eigenvalue that dilates distance by some factor $\rho > 1$ contributes $\log \rho$ to the entropy.

Returning to the case of an automorphism T_A of a solenoid corresponding to a matrix $A \in \mathrm{GL}_r(\mathbb{Q})$, write $\chi_A(t) = \det(A - tI_r) \in \mathbb{Q}[t]$ for the characteristic polynomial and let $s = s(A) \geqslant 1$ denote the least common multiple of the denominators of the coefficients of χ_A. Yuzvinskiĭ's formula states that

$$h(A) = \log s + \sum_{|\lambda|>1} \log|\lambda|, \tag{3.12}$$

where the sum is taken over the set of eigenvalues of A, with multiplicities. The second term in equation (3.12) accords exactly with our geometrical view of entropy, but the first term $\log s$ does not. It turns out that the two terms are on the same footing: both are sums of contributions $\log \rho$ corresponding to eigenvalues that dilate by a factor ρ, and we will prove in Theorems 3.16 and 3.17 that

$$h(A) = \sum_{p \leqslant \infty} \sum_{|\lambda_{j,p}|_p > 1} \log|\lambda_j|_p,$$

where $\{\lambda_{j,p}\}$ denotes the set of eigenvalues of A in a finite extension of $\mathbb{Q}_p$.

The first step is to show that a general solenoid can be simplified to the case $\widehat{\mathbb{Q}}^r$ without changing the entropy. Let $T : X \to X$ be an automorphism of a solenoid with $\widehat{X} \leqslant \mathbb{Q}^r$ (with r chosen to be minimal with this property); the automorphism $\widehat{T} : \widehat{X} \to \widehat{X}$ extends to an automorphism of $\mathbb{Q}^d$. If we write Σ for the solenoid $\mathbb{Q}^r$ and $T_{\mathbb{Q}}$ for the automorphism of Σ dual to this automorphism of $\mathbb{Q}^r$, then the injective map $\widehat{X} \to \mathbb{Q}^r$ dualizes to give a commutative diagram

$$\begin{array}{ccc} \Sigma & \xrightarrow{T_{\mathbb{Q}}} & \Sigma \\ \downarrow & & \downarrow \\ X & \xrightarrow[T]{} & X \end{array}$$

realizing T as a topological factor of $T_{\mathbb{Q}}$.

Lemma 3.15. *Let $T : X \to X$ be an automorphism of a solenoid. Then*

$$h_{\mathrm{top}}(T) = h_{\mathrm{top}}(T_{\mathbb{Q}}).$$

Proof. Recall that $T : X \to X$ is an automorphism, and $\widehat{X} \leqslant \mathbb{Q}^r$. For any $n \geqslant 1$ the subgroup $\frac{1}{n!}\widehat{X}$ is a $\widehat{T}$-invariant subgroup of $\mathbb{Q}^r$, this defines an increasing sequence of subgroups whose union is all of $\mathbb{Q}^r$. Thus $\mathbb{Q}^r$ is the direct limit

$$\widehat{X} \xrightarrow{\phi_1} \tfrac{1}{2!}\widehat{X} \xrightarrow{\phi_2} \cdots$$

where $\phi_n : \frac{1}{n!}\widehat{X} \to \frac{1}{(n+1)!}\widehat{X}$ is the map defined by $\phi_n(x) = x$. Writing Σ_n for the dual of $\frac{1}{n!}\widehat{X}$, this means that Σ is the projective limit

$$\Sigma_1 \xleftarrow{\widehat{\phi_1}} \Sigma_2 \xleftarrow{\widehat{\phi_2}} \cdots .$$

Equivalently,

$$\Sigma = \{(x_n) \in \Sigma_1 \times \Sigma_2 \times \cdots \mid x_n = \widehat{\phi_n} x_{n+1} \text{ for all } n \geqslant 1\}$$

with the topology induced from the product topology. The partition ξ_n into cosets of Σ_n is generating, so

$$h_{m_\Sigma}(T_{\mathbb{Q}}) = \lim_{n\to\infty} h_{m_{\Sigma_n}}(T_n)$$

where T_n is the induced map on Σ_n. On the other hand, as an abstract group each $\frac{1}{n!}\widehat{X}$ is isomorphic to $\widehat{X}$ (since they are all torsion-free groups), so $h_{m_{\Sigma_n}}(T_n) = h_{m_\Sigma}(T_{\mathbb{Q}})$ for all $n \geqslant 1$. By Theorem 3.10, this gives the same result for topological entropy. □

Thus we may assume that T is the map $T_A : \Sigma \to \Sigma$ where $\Sigma = \widehat{\mathbb{Q}}^r$ and $A \in \mathrm{GL}_r(\mathbb{Q})$. By duality, A acts as an automorphism of the r-dimensional solenoid Σ, and also as a uniformly continuous linear map on $\mathbb{Q}_p^r$ for $p \leqslant \infty$

(the infinite place giving $\mathbb{Q}^r_\infty = \mathbb{R}^r$). Write d_p for the maximum metric on $\mathbb{Q}^r_p$ for each $p \leqslant \infty$, and d for the metric on the various sets of the form $\mathbb{Q}_{\mathbb{A}}(P)^r$ arising in the proof. The entropy of T_A, dual to the action of A on the vector space $\mathbb{Q}^r$ over the global field $\mathbb{Q}$, decomposes into a sum of local contributions corresponding to the places of $\mathbb{Q}$.

Theorem 3.16. $h_{\mathrm{top}}(T_A) = \sum_{p \leqslant \infty} h_{\mathsf{d}_p}(A \text{ on } \mathbb{Q}^r_p)$.

Proof. The adele ring $\mathbb{Q}^r_{\mathbb{A}}$ contains $\mathbb{Q}^r$ as a discrete subring via the map δ (by Theorem B.10), so we may view it as a $\mathbb{Q}$-vector space and thus extend the action of A on $\mathbb{Q}^r$ to a uniformly continuous map on $\mathbb{Q}^r_{\mathbb{A}}$ by defining

$$(Ax)_p = A(x_p)$$

for $x \in \mathbb{Q}^r_p$. Under this action the embedded copy of $\mathbb{Q}^r$ is invariant, so by Theorem B.12, the induced action of A on $\mathbb{Q}^r_{\mathbb{A}}/\delta(\mathbb{Q}^r)$ is isomorphic to the action of $T_{\mathbb{Q}}$ on Σ. The quotient map

$$\mathbb{Q}^r_{\mathbb{A}} \to \mathbb{Q}^r_{\mathbb{A}}/\delta(\mathbb{Q}^r)$$

is a local isometry (since $\delta(\mathbb{Q}^r)$ is a discrete subgroup), so

$$h(T_{\mathbb{Q}}) = h(A \text{ on } \mathbb{Q}^r_{\mathbb{A}})$$

by Proposition 3.8.

Just as in Section 3.3, we use Haar measure to compute entropy via the decay of volume of Bowen balls.

Both A and A^{-1} have entries in $\mathbb{Z}_p$ for all but finitely many p; let P_A be the set of primes p for which some entry of A or of A^{-1} is not in $\mathbb{Z}_p$, together with ∞. Thus $A \in \mathrm{GL}_r(\mathbb{Z}_p)$ for any $p \notin P$, so

$$\mathbb{Q}_{\mathbb{A}}(P)^r = \prod_{p \in P} \mathbb{Q}^r_p \times \prod_{p \notin P} \mathbb{Z}_p$$

(as in Section B.2) is an A-invariant neighborhood of the identity. By Theorem 3.10, it follows that

$$h_{\mathsf{d}}(A \text{ on } \mathbb{Q}^r_{\mathbb{A}}) = h(A \text{ on } \mathbb{Q}_{\mathbb{A}}(P)^r).$$

Now the second factor in $\mathbb{Q}_{\mathbb{A}}(P)^r = \prod_{p \in P} \mathbb{Q}^r_p \times \prod_{p \notin P} \mathbb{Z}^r_p$ is compact, so

$$h_{\mathsf{d}}\left(A \text{ on } \mathbb{Q}_{\mathbb{A}}(P)^r\right) = h_{\mathsf{d}}\left(A \text{ on } \prod_{p \in P} \mathbb{Q}^r_p\right) + h_{\mathsf{d}}\left(A \text{ on } \prod_{p \notin P} \mathbb{Z}^r_p\right). \qquad (3.13)$$

by Lemma 3.7. If F is any finite set of primes in the complement Q of P then, for any $m \geqslant 1$,

$$\prod_{p\in F} p^m\mathbb{Z}_p^r \times \prod_{p\in Q\setminus F} \mathbb{Z}_p^r$$

is an A-invariant neighborhood of the identity since $A \in \mathrm{GL}_r(\mathbb{Z}_p)$ for $p \notin P$. These neighborhoods form a basis for the topology, so $h(A \text{ on } \prod_{p\notin P}\mathbb{Z}_p^r) = 0$. We deduce that

$$h_{\mathsf{d}}\left(A \text{ on } \mathbb{Q}_{\mathbb{A}}(P)^r\right) = h_{\mathsf{d}}\left(A \text{ on } \prod_{p\in P}\mathbb{Q}_p^r\right).$$

Now the Haar measure $m_{\mathbb{Q}_p}$ on $\mathbb{Q}_p$, normalized to have $m_{\mathbb{Q}_p}(\mathbb{Z}_p) = 1$ for all $p < \infty$, and the Haar measure

$$m_{\mathbb{Q}_{\mathbb{A}}} = \prod_{p\leqslant\infty} m_{\mathbb{Q}_p}$$

are A-homogeneous, and we will show in the proof of Theorem 3.17 that

$$-\frac{1}{n}\log m^r_{\mathbb{Q}_p}\left(\bigcap_{k=0}^{n-1} A^{-k}(B^r)\right) \longrightarrow h(A \text{ on } \mathbb{Q}_p^r) \tag{3.14}$$

as $n \to \infty$ for $p < \infty$ (in particular, showing the convergence); for the real case $p = \infty$ this is shown in the proof of Theorem 3.12. By Lemma 3.11, it follows that

$$h_{\mathsf{d}}\left(A \text{ on } \mathbb{Q}_{\mathbb{A}}(P)^r\right) = \sum_{p\in P} h_{\mathsf{d}_p}(A \text{ on } \mathbb{Q}_p^r).$$

For $p \notin P$, the fact that $p^m\mathbb{Z}_p$ forms a basis of A-invariant open neighborhoods of the identity shows that $h_{\mathsf{d}_p}(A \text{ on } \mathbb{Q}_p^r) = 0$, so this completes the proof of Theorem 3.16 by equation (3.13). □

We now turn to the calculation of the local entropies appearing in Theorem 3.16. For $p = \infty$ we have done this already, so it is sufficient to consider the case of a finite prime.

Theorem 3.17. *Assume that $p < \infty$. If A has eigenvalues $\lambda_1, \ldots, \lambda_r$ in a finite extension of $\mathbb{Q}_p$, then*

$$h_{\mathsf{d}_p}(A \text{ on } \mathbb{Q}_p^r) = \sum_{|\lambda_j|_p>1} \log|\lambda_j|_p,$$

where eigenvalues are counted with multiplicity and $|\cdot|_p$ denotes the unique extension of the p-adic norm to the splitting field of the characteristic polynomial of A. Moreover, there is convergence in equation (3.14).

The proof below follows exactly the lines of the case $p = \infty$ in Theorem 3.12, but the ultrametric inequality in $\mathbb{Q}_p$ makes the analysis considerably easier. The entire argument is visible in Figure 3.1.

Outline Proof of Theorem 3.17. Let K be a finite extension of $\mathbb{Q}_p$ containing all zeros of the characteristic polynomial χ_A of A, and set $d = [K : \mathbb{Q}_p]$. Then $\mathbb{Q}_p^r \otimes_{\mathbb{Q}_p} K \cong K^r$, and A extends to the map $A \otimes 1_K$ acting on K^r. Since K is a d-dimensional vector space over $\mathbb{Q}_p$, and $A \in \mathrm{GL}_r(\mathbb{Q}) \leqslant \mathrm{GL}_r(\mathbb{Q}_p)$, the map $A \otimes 1_K$ is isomorphic to the direct sum of d copies of the map A acting on $\mathbb{Q}_p^r$. Thus $h_{\mathtt{d}_p}(A \otimes 1_K \text{ on } K^r) = dh_{\mathtt{d}_p}(A \text{ on } \mathbb{Q}_p^r)$.

Since K contains the eigenvalues of $A \otimes 1_K$, we can proceed just as in the final part of the proof of Theorem 3.12 and put $A \otimes 1_K$ into its Jordan form

$$A \otimes 1_K \cong \begin{bmatrix} J(\lambda_1, d_1) & & & \\ & J(\lambda_2, d_2) & & \\ & & \ddots & \\ & & & J(\lambda_k, d_k) \end{bmatrix}$$

where $J(\lambda_i, d_i)$ denotes the Jordan block of size d_i corresponding to λ_i. Once convergence is established for the decay of Bowen balls in each block, the entropy sums over the blocks so it is sufficient to prove the result for one block $J = J(\lambda, m)$ acting on K^m equipped with the maximum norm. We may assume without loss that $|\lambda|_p > 1$, since otherwise both sides are 0, and the calculation used in the case of $\mathbb{R}$ or $\mathbb{C}$ shows that there is convergence and the formula required. Finally,

$$\begin{aligned} h_{\mathtt{d}_p}(A \text{ on } \mathbb{Q}_p^r) &= \frac{1}{r} h(A \otimes 1_K \text{ on } K^r) \\ &= \frac{1}{r} \sum_{i=1}^{k} h(J(\lambda_i, d_i) \text{ on } K^{d_i}) \\ &= \frac{1}{r} \sum_{i=1}^{k} r d_i \log^+ |\lambda_i|_p = \sum_{|\lambda_j|_p > 1} \log^+ |\lambda_j|_p. \end{aligned}$$

□

Thus the entropy of an automorphism of a solenoid is a sum over local contributions. The way in which these contributions fit together is illustrated by some examples.

Example 3.18. If $A \in \mathrm{GL}_d(\mathbb{Z})$, then for any $p < \infty$ we have $|\lambda_{i,p}|_p = 1$ for all of the eigenvalues $\lambda_{i,p}$ of A in an extension of $\mathbb{Q}_p$[24]. Thus all the entropy of the automorphism of Σ^d induced by A comes from the infinite place, with all the finite contributions being zero.

Example 3.19. An opposite extreme to Example 3.18 is given by an example used by Lind [84] to show that a general exponential rate of recurrence phenomena for ergodic group automorphisms may be driven entirely by p-adic hyperbolicity. Let

$$A = \begin{bmatrix} 0 & -1 \\ 1 & \frac{6}{5} \end{bmatrix},$$

with characteristic polynomial $\chi_A(t) = t^2 - \frac{6}{5}t + 1$. The complex eigenvalues $\frac{3}{5} \pm \frac{4}{5}\mathrm{i}$ of A have modulus 1, so the entropy contribution from the infinite place is zero. The 5-adic eigenvalues λ_1, λ_2 satisfy $|\lambda_1\lambda_2|_5 = |1|_5 = 1$ and $|\lambda_1 + \lambda_2|_5 = |\frac{6}{5}|_5 = 5$, so $|\lambda_1|_5 = 5$ and $|\lambda_2|_5 = \frac{1}{5}$. For any finite $p \neq 5$ the p-adic eigenvalues λ_1, λ_2 satisfy $|\lambda_1\lambda_2|_p = 1$ and

$$|\lambda_1 + \lambda_2|_p = |\tfrac{6}{5}|_p \begin{cases} < 1 & \text{if } p \in \{2, 3\}; \\ = 1 & \text{if not.} \end{cases}$$

It follows that $|\lambda_1|_p = |\lambda_2|_p = 1$, since if one of $|\lambda_1|_p$ or $|\lambda_2|_p$ exceeds 1, then

$$|\lambda_1 + \lambda_2|_p = |\lambda_1 + \lambda_1^{-1}|_p > 1.$$

Thus

$$h_{\mathrm{top}}(A \text{ on } \Sigma^2) = h_{\mathrm{top}}(A \text{ on } \mathbb{Q}_5^2) = \log 5,$$

and the only positive contribution comes from the 5-adic place.

Example 3.20. In general we expect there to be a mixture of infinite and finite contributions to entropy, and this may already be seen in the case of a one-dimensional solenoid X. Here an automorphism is defined by a matrix $[\frac{a}{b}]$ in $\mathrm{GL}_1(\mathbb{Q}) = \mathbb{Q}^\times$, written as a rational $\frac{a}{b}$ in lowest terms, and Theorems 3.16 and 3.17 show that

$$h_{\mathrm{top}}([\tfrac{a}{b}] \text{ on } X) = \textstyle\sum_{p \leqslant \infty} \log^+ |\frac{a}{b}|_p = \log \max\{|a|, |b|\}, \tag{3.15}$$

with contributions only coming from the infinite place and those $p < \infty$ dividing b. This recovers a formula due to Abramov [1]. For example,

$$h_{\mathrm{top}}([\tfrac{3}{2}] \text{ on } \Sigma) = \log^+ |\tfrac{3}{2}| + \log^+ |\tfrac{3}{2}|_2 + \log^+ |\tfrac{3}{2}|_3 = \log \tfrac{3}{2} + \log 2 + 0 = \log 3.$$

On the other hand, Lemma 2.19 shows that $h_{\mathrm{top}}([\frac{2}{3}] \text{ on } \Sigma) = \log 3$ also, and this arises entirely from the 3-adic contribution:

$$h_{\mathrm{top}}([\tfrac{2}{3}] \text{ on } \Sigma) = \log^+ |\tfrac{2}{3}| + \log^+ |\tfrac{2}{3}|_2 + \log^+ |\tfrac{2}{3}|_3 = 0 + 0 + \log 3 = \log 3.$$

Finally, we show how Theorems 3.16 and 3.17 together show Yuzvinskiĭ's rather cryptic formula in equation (3.12). As before, we let $A \in \mathrm{GL}_d(\mathbb{Q})$ be a rational matrix with eigenvalues $\lambda_1, \ldots, \lambda_d \in \mathbb{C}$, and let s be the least common multiple of the denominators of the coefficients of the characteristic polynomial $\chi_A(t)$.

Lemma 3.21 (Yuzvinskiĭ). *Let $A \in \mathrm{GL}_d(\mathbb{Q})$. Then*

$$h(A \text{ on } \Sigma^d) = \log s + \sum_{|\lambda_j| > 1} \log |\lambda_j|$$

where the sum is taken over the eigenvalues of A.

PROOF. Let $\chi_A(t) = t^d + a_1 t^{d-1} + \cdots + a_d$. If $|s|_p = p^{-e}$, then

$$p^e = \max\{|a_1|_p, \ldots, |a_d|_p, 1\}.$$

We claim that

$$h_{\text{top}}(A \text{ on } \mathbb{Q}_p^d) = \log p^e. \tag{3.16}$$

Theorem 3.17 then shows that $\log s$ the sum over finite primes p of the p-adic contributions to the topological entropy, proving Lemma 3.21. All that remains is to show equation (3.16). The characteristic polynomial factorizes as

$$\chi_A(t) = \prod_{j=1}^{d} (t - \lambda_j)$$

over a finite extension of $\mathbb{Q}_p$, and we may arrange the eigenvalues so that

$$|\lambda_1|_p \geqslant |\lambda_2|_p \geqslant \cdots \geqslant |\lambda_m|_p > 1 \geqslant |\lambda_{m+1}|_p \geqslant \cdots \geqslant |\lambda_d|_p.$$

If $|\lambda_j|_p \leqslant 1$ for all j, then $e = 0$ and $h(A \text{ on } \mathbb{Q}_p^d) = 0$ also. Thus we may suppose that $|\lambda_1|_p > 1$. Using the ultrametric inequality, we have

$$\begin{aligned} |a_m|_p &= \left| \sum_{i_1 < \cdots < i_m} \lambda_{i_1} \ldots \lambda_{i_m} \right|_p \\ &= |\lambda_1 \ldots \lambda_m + \text{ terms smaller in } p\text{-adic norm}|_p \\ &= |\lambda_1 \ldots \lambda_m|_p, \end{aligned}$$

and

$$|a_j|_p \leqslant |a_m|_p$$

similarly. Thus

$$p^e = \max_{1 \leqslant j \leqslant d} \{|a_j|_p\} = \prod_{|\lambda_j^{(p)}|_p > 1} |\lambda_j^{(p)}|_p = h_{\text{top}}(A \text{ on } \mathbb{Q}_p^d),$$

completing the proof. □

Exercises for Section 3.3

Exercise 3.3.1. Show that Lemma 2.19 is false without the assumption that X is compact.

Exercise 3.3.2. Strengthen the inequality for entropy of topological factors from Exercise 2.2.2 as follows. If $T_i : (X_i, \mathsf{d}_i) \to (X_i, \mathsf{d}_i)$ for $i = 1, 2$ are continuous maps of compact metric spaces, and $\pi : X_1 \to X_2$ is a continuous surjective factor map, then

$$h_{\mathsf{d}_1}(T_1) \leqslant h_{\mathsf{d}_2}(T_2) + \sup_{x \in X_2} h_{\mathsf{d}_1}(T_1, \pi^{-1}(x)).$$

Exercise 3.3.3. Prove the second equality in equation (3.15), completing the proof of Abramov's formula.

Exercise 3.3.4. Show that the entire argument in Section 3.3.2 extends to an automorphism dual to the action of $A \in \mathrm{GL}_d(K)$ on K^d for any $\mathbb{A}$-field K.

3.3.3 Flows on Compact Homogeneous Surfaces

In this section* G will denote a unimodular Lie group. Recall that a *uniform lattice*† is a discrete subgroup $\Gamma \leqslant G$ for which $X = \Gamma\backslash G$ is compact. Recall (see, for example, [38, Sect. 9.3.3]) that it is possible to define a left-invariant metric $\mathsf{d}_G(\cdot,\cdot)$ on G which then via

$$\mathsf{d}_X(\Gamma g_1, \Gamma g_2) = \min_{\gamma\in\Gamma} \mathsf{d}_G(g_1, \gamma g_2)$$

defines a metric on any quotient $X = \Gamma\backslash G$ by a discrete subgroup $\Gamma \leqslant G$. Moreover, for every compact subset $K \subseteq X$ there exists some $r > 0$ (called an *injectivity radius*) with the property that

$$B_r^G \ni h \longmapsto xh \in B_r^X(x)$$

is an isometry for every $x \in K$. Hence, as we are assuming that $X = \Gamma\backslash G$ is compact, this implies the assumptions regarding the metric in Proposition 3.8 and the canonical map $\pi : (G, \mathsf{d}_G) \to (X, \mathsf{d}_X)$ defined by $\pi(g) = \Gamma g$.

Now fix some $a \in G$ and define

$$T_G : g \longmapsto ga^{-1}$$

for $g \in G$, respectively

$$T_X : x \longmapsto xa^{-1}$$

for $x \in X$. Then we also have the commutative diagram in Proposition 3.8, which implies that $h_{\mathsf{d}_G}(T_G) = h_{\mathsf{d}_X}(T_X)$. These quantities can be calculated as follows.

* In this section we follow Bowen [14] closely. We would like to point out that the material in Sections 3.3.1 and 3.3.2 gives an easy route to a 'formula' for the entropy of certain maps at the expense of masking the detailed dynamics. On the other hand, the material in Section 4.5 will give greater insights (and in particular will give a characterization of Lebesgue measure as the unique maximal measure). For that reason, we will revisit this kind of calculation again by generalizing the approach taken in Section 4.5.

† A lattice is a discrete subgroup $\Gamma \leqslant G$ for which the quotient space $\Gamma\backslash G$ has finite volume. Unfortunately, Proposition 3.8 does not apply in this setting, hence the need to restrict to compact quotients.

Theorem 3.22. *Let $X = \Gamma\backslash G$ be a compact quotient of a unimodular* Lie group by a discrete subgroup. Let $a \in G$, and define $T_X(x) = xa^{-1}$ for $x \in X$. Then the topological entropy of T_X is given by*

$$h_{\text{top}}(T_X) = k(m_G, T_G) = k(m_{\mathfrak{g}}, \text{Ad}_a) = \sum_{i=1}^{\dim(G)} \log^+ |\lambda_i|,$$

where $\lambda_1, \ldots, \lambda_{\dim(G)}$ are the eigenvalues (listed with algebraic multiplicity) of the linear map Ad_a on the Lie algebra $\mathfrak{g}$ of G, and $m_{\mathfrak{g}}$ is the Lebesgue measure on $\mathfrak{g} \cong \mathbb{R}^{\dim(G)}$.

Proof. By the discussion before the statement of the theorem, we have

$$h_{\mathsf{d}_X}(T_X) = h_{\mathsf{d}_G}(T_G).$$

Let $\theta_a(g) = aga^{-1}$ denote conjugation by a as a map on G. Since d_G is invariant under left multiplication, it follows that

$$\mathsf{d}_G\left(\theta_a^k(y), \theta_a^k(x)\right) < \varepsilon$$

if and only if

$$\mathsf{d}_G\left(T_G^k(y), T_G^k(x)\right) < \varepsilon,$$

for any $k \in \mathbb{Z}$. This implies that $D_n(x, \varepsilon, T_G) = D_n(x, \varepsilon, \theta_a)$ is the Bowen ball for T_G and for θ_a. Similarly, $D_n(hx, \varepsilon, T_G) = hD_n(x, \varepsilon, T_G)$ for any $h \in G$, which implies that the bi-invariant Haar measure m_G is homogeneous (in the sense of Definition 3.9) both for T_G and for θ_a. To summarize, we have shown that so far that

$$h_{\text{top}}(T_X) = h_{\mathsf{d}_X}(T_X) = h_{\mathsf{d}_G}(T_G) = k(m_G, \theta_a).$$

On the other hand $\text{Ad}_a : \mathfrak{g} \to \mathfrak{g}$ is linear, and so Theorem 3.12 shows that

$$h_{\mathsf{d}_{\mathfrak{g}}}(\text{Ad}_a) = k(m_{\mathfrak{g}}, \text{Ad}_a) = \sum_{\lambda} \log^+ |\lambda|.$$

Thus it remains only to show that $k(m_G, \theta_a) = k(m_{\mathfrak{g}}, \text{Ad}_a)$, which will follow by analyzing how the (measures of the) Bowen balls $D_n(e, \varepsilon, \theta_a)$ at the identity in the Lie group, and $D_n(0, \delta, \text{Ad}_a)$ at 0 in the Lie algebra relate to each other.

Recall that

$$\begin{aligned} \exp : \mathfrak{g} &\longrightarrow G \\ v &\longmapsto \exp(v) \end{aligned}$$

is a diffeomorphism when restricted to be a map from some open neighborhood of $0 \in \mathfrak{g}$ to some open neighborhood of $e \in G$. Moreover,

* The existence of a lattice in G implies that G is unimodular (see [38, Prop. 9.20]).

$$\exp(\mathrm{Ad}_a(v)) = \theta_a(\exp(v))$$

for any $v \in \mathfrak{g}$. This clearly implies that for every $\varepsilon > 0$ there is some $\delta > 0$ with

$$\exp\left(D_n(0, \delta, \mathrm{Ad}_a)\right) \subseteq D_n(e, \varepsilon, \theta_a), \tag{3.17}$$

and also that for every $\delta > 0$ there exists an $\varepsilon > 0$ with

$$\exp\left(D_n(0, \delta, \mathrm{Ad}_a)\right) \supseteq D_n(e, \varepsilon, \theta_a). \tag{3.18}$$

Finally, recall that the Haar measure m_G is a smooth measure. In particular there exists a neighborhood $B^{\mathfrak{g}}_{\delta_0}$ of $0 \in \mathfrak{g}$, and some constant $C \geqslant 1$ such that

$$\frac{m_G(\exp(B))}{m_{\mathfrak{g}}(B)} \in [\tfrac{1}{C}, C] \tag{3.19}$$

for any Borel subset $B \subseteq B^{\mathfrak{g}}_{\delta_0}$. Now equation (3.17)–(3.19) and the definition in equation (3.7) together imply that $k(m_G, \theta_a) = k(m_{\mathfrak{g}}, \mathrm{Ad}_a)$ as required. □

Notes to Chapter 3

[18](Page 70) The following example to illustrate this is taken from Walters [140, Sect. 7.2]. Let d_1 be the usual metric on $X = (0, \infty)$; define another metric d_2 to coincide with d_1 on $(1, 2]$ and so as to make the map $T : X \to X$ defined by $T(x) = 2x$ an isometry for d_2. This is possible since the orbit of $(1, 2]$ under T partitions the whole space,

$$(0, \infty) = \cdots \sqcup (\tfrac{1}{4}, \tfrac{1}{2}] \sqcup (\tfrac{1}{2}, 1] \sqcup (1, 2] \sqcup \cdots .$$

Notice that d_1 and d_2 define the same topology on X. Since $T : (X, \mathsf{d}_2) \to (X, \mathsf{d}_2)$ is an isometry, $h_{\mathsf{d}_2}(T) = 0$. On the other hand,

$$\max_{0 \leqslant j \leqslant n-1} \{\mathsf{d}_1(T^j x, T^j y)\} = 2^{n-1}|x - y|,$$

so the least cardinality of an $(n, \frac{1}{2^k})$-spanning set is $2^{n-1}2^k + 1$. This shows that $h_{\mathsf{d}_1}(T) \geqslant \log 2$, showing that the topological entropy depends on the uniform equivalence class of the metric, not just the equivalence class.

[19](Page 72) Corollary 3.6 is shown by Adler, Konheim and McAndrew [3]; Lemma 2.20 is shown by Goodwyn [56] (both have a purely topological argument, without the use of a metric: as pointed out by Goodwyn, even for compact topological spaces it is necessary to assume that the spaces are Hausdorff in order to know that $N(\mathscr{U} \times \mathscr{V}) = N(\mathscr{U}) \times N(\mathscr{V})$). Lemmas 3.5 and 3.7 are taken from Bowen [14]; the possible non-additivity over products is missed in [14] and corrected in [15]. The only obstacle to additivity is that we cannot pass from the inequality $a_n \geqslant b_n + c_n$ to the inequality $\limsup_{n\to\infty} a_n \geqslant \limsup_{n\to\infty} b_n + \limsup_{n\to\infty} c_n$ unless we know that there is convergence (as assumed in the second part of Lemma 3.7 and in Lemma 3.11(1), or we know that $b_n = c_n$ for all $n \geqslant 1$ as in Lemma 3.11(2).)

[20](Page 78) This idea was developed by Bowen [14], [15]; he used it to compute the topological entropy of affine maps of Lie groups and other homogeneous spaces,

and to show that Haar measure is maximal for affine maps. Much of the material in Section 3.3 comes from [14]. Theorem 3.13 (the generalization of Theorem 4.24 to automorphisms of the r-torus) was shown for $r = 2$ by Sinaĭ [131]; the general case was stated in [131] and in a paper of Genis [48]. Arov [7] gave a proof as part of his calculation of the entropy of endomorphisms of solenoids (these are generalizations of the torus). Berg [8] gave an independent proof using different methods. Finally Yuzvinskiĭ [148] computed the entropy of any compact group endomorphism. For modern treatments, see Walters [140, Chap. 7] for toral automorphisms; Lind and Ward [88] for automorphisms of the solenoid, as discussed in Section 3.3.2.

[21] (Page 80) There are several ways to do this; the path chosen here is done so in order to relate the calculation for each block directly to a real part of the space on which the matrix acts, rather than a different complexified space.

[22] (Page 83) A simple example of a solenoid is a torus; in [38, Exercise 2.1.9] the solenoid dual to $\mathbb{Z}[\frac{1}{2}]$ is constructed. The results in this section formally subsume Theorem 3.12, but the proofs are not a generalization of the proof of Theorem 3.12; indeed we will use that theorem to prove the more general case. To see how diverse solenoids are, notice that for any subset S of the set of rational primes, the ring of S-integers R_S (see Section B.2) is a subgroup of $\mathbb{Q}$, and if $S \neq S'$ then R_S and $R_{S'}$ are not isomorphic. This shows there are uncountably many different one-dimensional solenoids; an easy calculation confirms that these may be used to find uncountably many topologically distinct algebraic dynamical systems with the same topological entropy [142].

[23] (Page 83) We follow Lind and Ward [88], [141] closely here. The possibility of computing entropy for solenoidal automorphisms using p-adic entropy contributions was suggested by Lind in 1980, and ultimately goes back to a suggestion of Furstenberg.

[24] (Page 87) This is most easily seen using the Newton polygon of the characteristic polynomial of A (see Koblitz [75] for example); the proof of Lemma 3.21 also shows this in a more general setting.

4

Conditional Measure-Theoretic Entropy

Infinite sub-σ-algebras and infinite partitions behave differently, and in particular the correspondence between partitions and sub-σ-algebras only goes in one direction. If ξ is a countably infinite partition, then $\sigma(\xi)$ is an uncountable σ-algebra. However, σ-algebras of the form $\sigma(\xi)$ where ξ is a countable partition are rather special, and should not be confused with the much larger class of countably-generated σ-algebras. One way to describe the difference between the two classes of σ-algebras is via the properties of their atoms: In general a countably-generated σ-algebra may have uncountably many atoms, each of zero measure, while the σ-algebra $\sigma(\xi)$ will only have countably many atoms, namely the elements of ξ.

The basic entropy theory from Chapter 1 will become a more powerful and flexible tool after we extend the theory from partitions to σ-algebras.

4.1 Conditional Entropy

We recall from [38, Sect. 5.3] in the case of a countable partition) the construction of conditional measures: if $\mathscr{A} \subseteq \mathscr{B}$ is a sub-σ-algebra in the probability space $(X, \mathscr{B}, \mu)$, then there exists a family of conditional measures $\{\mu_x^{\mathscr{A}} \mid x \in X\}$ with

$$\mu = \int \mu_x^{\mathscr{A}} \, \mathrm{d}\mu.$$

In a sense which we will make precise below, $\mu_x^{\mathscr{A}}$ describes μ restricted to the atom $[x]_{\mathscr{A}}$ in a way that makes sense even if the atom is a null set. If $\mathscr{A}$ is countably generated, then the atom of $\mathscr{A}$ containing x is the set

$$[x]_{\mathscr{A}} = \bigcap_{x \in A \in \mathscr{A}} A,$$

which is the smallest element of $\mathscr{A}$ containing x. If $\mathscr{C} = \sigma(\xi)$ for a partition ξ, then the atoms are just the elements of the partition. These conditional mea-

sures exist if $(X, \mathscr{B}, \mu)$ is a Borel probability spaces*. Also, if $\mathscr{C}$ is a countably-generated σ-algebra, then X decomposes into atoms $[x]_{\mathscr{C}}$ for $x \in X$.

We recall a simple result extending the relation between conditional measures and conditional expectations from [38, Chap. 5].

Theorem 4.1. *Let $(X, \mathscr{B}, \mu)$ be a Borel probability space, and $\mathscr{A} \subseteq \mathscr{B}$ a σ-algebra. Then there exists an $\mathscr{A}$-measurable conull set $X' \subseteq X$ and a system $\{\mu_x^{\mathscr{A}} \mid x \in X'\}$ of measures on X, referred to as* conditional measures, *with the following properties.*

(1) *$\mu_x^{\mathscr{A}}$ is a probability measure on X with*

$$E(f\,|\,\mathscr{A})(x) = \int f(y)\,\mathrm{d}\mu_x^{\mathscr{A}}(y) \tag{4.1}$$

almost everywhere for all $f \in \mathscr{L}^1(X, \mathscr{B}, \mu)$. In other words, for any function† $f \in \mathscr{L}^1(X, \mathscr{B}, \mu)$ we have that $\int f(y)\,\mathrm{d}\mu_x^{\mathscr{A}}(y)$ exists for all x belonging to a conull set in $\mathscr{A}$, that on this set

$$x \mapsto \int f(y)\,\mathrm{d}\mu_x^{\mathscr{A}}(y)$$

depends $\mathscr{A}$-measurably on x, and that

$$\int_A \int f(y)\,\mathrm{d}\mu_x^{\mathscr{A}}(y)\,\mathrm{d}\mu(x) = \int_A f\,\mathrm{d}\mu$$

for all $A \in \mathscr{A}$.

(2) *If $\mathscr{A}$ is countably-generated, then $\mu_x^{\mathscr{A}}([x]_{\mathscr{A}}) = 1$ for all $x \in X'$, where*

$$[x]_{\mathscr{A}} = \bigcap_{x \in A \in \mathscr{A}} A$$

is the atom of $\mathscr{A}$ *containing x; moreover $\mu_x^{\mathscr{A}} = \mu_y^{\mathscr{A}}$ for $x, y \in X'$ whenever $[x]_{\mathscr{A}} = [y]_{\mathscr{A}}$.*

(3) *Property* (1) *uniquely determines $\mu_x^{\mathscr{A}}$ for a.e. $x \in X$. In fact, property* (1) *for a dense countable set of functions in $C(\overline{X})$ uniquely determines $\mu_x^{\mathscr{A}}$ for a.e. $x \in X$.*

(4) *If $\widetilde{\mathscr{A}}$ is any σ-algebra with $\mathscr{A} \underset{\mu}{=} \widetilde{\mathscr{A}}$, then $\mu_x^{\mathscr{A}} = \mu_x^{\widetilde{\mathscr{A}}}$ almost everywhere.*

* A Borel probability space is here taken to mean a dense Borel subset of a compact metric space $\overline{X}$, with a probability measure μ defined on the restriction of the Borel σ-algebra $\mathscr{B}$ to X. This simplification could be avoided in most statements by using the conditional expectation instead of conditional measures.

† Notice that we are forced to work with genuine functions in $\mathscr{L}^1$ in order to ensure that the right-hand side of equation (5.3) is defined. As we said before, $\mu_x^{\mathscr{A}}$ may be singular to μ.

This is simply [38, Th. 5.14], and it is proved in [38, Sect. 5.3]. We also recall from that section the following example.

Example 4.2. Let $X = [0,1]^2$ and $\mathscr{A} = \mathscr{B} \times \{\varnothing, [0,1]\}$. Theorem 4.1 says that any Borel probability measure μ on X can be decomposed into vertical components in the following sense: the conditional measures $\mu^{\mathscr{A}}_{(x_1,x_2)}$ are defined on the line segments $\{x_1\} \times [0,1]$, and these sets are precisely the atoms of $\mathscr{A}$. Moreover,

$$\mu(B) = \int_X \mu^{\mathscr{A}}_{(x_1,x_2)}(B)\, \mathrm{d}\mu(x_1,x_2). \tag{4.2}$$

Here $\mu^{\mathscr{A}}_{(x_1,x_2)} = \nu_{x_1}$ does not depend on x_2, so equation (4.2) may be written as

$$\mu(B) = \int_{[0,1]} \nu_{x_1}(B)\, \mathrm{d}\overline{\mu}(x_1)$$

where $\overline{\mu} = \pi_*\mu$ is the measure on $[0,1]$ obtained by the projection

$$\begin{aligned} \pi : [0,1]^2 &\longrightarrow [0,1] \\ (x_1,x_2) &\longmapsto x_1. \end{aligned}$$

Example 4.2 clearly shows that we should think of Theorem 4.1 as a generalization of Fubini's theorem.

4.1.1 Conditional Information Functions

Definition 4.3. *Let $(X, \mathscr{B}, \mu)$ be a Borel probability space with $\mathscr{A}, \mathscr{C} \subseteq \mathscr{B}$ sub σ-algebras. The* information function of $\mathscr{C}$ given (the information of) $\mathscr{A}$ *with respect to μ is defined to be*

$$I_\mu\left(\mathscr{C}\middle|\mathscr{A}\right)(x) = -\log \mu^{\mathscr{A}}_x\left([x]_{\mathscr{C}}\right).$$

Moreover, the conditional entropy of $\mathscr{C}$ given $\mathscr{A}$,

$$H_\mu\left(\mathscr{C}\middle|\mathscr{A}\right) = \int I_\mu\left(\mathscr{C}\middle|\mathscr{A}\right)(x)\, \mathrm{d}\mu(x),$$

is defined to be the average of the information.

We will see later that $x \mapsto I_\mu\left(\mathscr{C}\middle|\mathscr{A}\right)(x)$ is measurable (see Proposition 4.5(2)). Notice that

$$I_\mu\left(\mathscr{C}\middle|\mathscr{A}\right) = I_\mu\left(\mathscr{A} \vee \mathscr{C}\middle|\mathscr{A}\right)$$

almost everywhere, since $[x]_{\mathscr{A}\vee\mathscr{C}} = [x]_{\mathscr{A}} \cap [x]_{\mathscr{C}}$ and $\mu^{\mathscr{A}}_x([x]_{\mathscr{A}}) = 1$ by Theorem 4.1.

If $\mathscr{C} \underset{\mu}{=} \mathscr{C}'$, then there is a null set N such that

$$[x]_{\mathscr{C}} \backslash N = [x]_{\mathscr{C}'} \backslash N$$

for all $x \in X$. Since $\mu_x^{\mathscr{A}}(N) = 0$ for almost every x by Theorem 4.1, we deduce that

$$I_\mu\left(\mathscr{C}\middle|\mathscr{A}\right) = I_\mu\left(\mathscr{C}'\middle|\mathscr{A}\right)$$

almost everywhere. Similarly, if $\mathscr{A} \underset{\mu}{=} \mathscr{A}'$ then $\mu_x^{\mathscr{A}} = \mu_x^{\mathscr{A}'}$ almost everywhere, and once again $I_\mu(\mathscr{C}\,|\,\mathscr{A}) = I_\mu(\mathscr{C}\,|\,\mathscr{A}')$ almost everywhere. Finally, notice that if $\mathscr{N} = \{X, \varnothing\}$ is the trivial σ-algebra, then

$$I_\mu(\mathscr{C}\big|\mathscr{N})(x) = I_\mu(\mathscr{C})(x) = -\log \mu([x]_{\mathscr{C}})$$

and

$$H_\mu(\mathscr{C}\big|\mathscr{N}) = H_\mu(\mathscr{C}) = \int I_\mu(\mathscr{C})\,\mathrm{d}\mu$$

is infinite unless $\mathscr{C} = \sigma(\xi)$ is the σ-algebra generated by a countable partition with finite entropy.

Just as in the case of partitions discussed on page 9, the information function of $\mathscr{C}$ given $\mathscr{A}$ at x is a measure of how much additional information is revealed by finding out which atom $[\cdot]_{\mathscr{C}}$ contains x starting from the knowledge of the atom $[x]_{\mathscr{A}}$. This informal description may help to motivate the following discussion, and the reader may find it helpful to find similar informal descriptions of the technical statements below.

Example 4.4. If $\mathscr{C} = \sigma(\xi)$ and $\mathscr{A} = \sigma(\eta)$ for countable partitions ξ and η, then the conditional measure $\mu_x^{\mathscr{A}} = \frac{1}{\mu([x]_\xi)}\mu|_{[x]_\xi}$ and so

$$I_\mu\left(\sigma(\xi)\middle|\sigma(\eta)\right)(x) = -\log\frac{\mu(P\cap Q)}{\mu(Q)} \quad \text{for } x \in P \in \xi, x \in Q \in \eta,$$

and

$$H_\mu\left(\sigma(\xi)\middle|\sigma(\eta)\right) = -\sum_{\substack{P\in\xi,\\ Q\in\eta}} \mu(P\cap Q)\log\frac{\mu(P\cap Q)}{\mu(Q)}.$$

Thus in this case the definition of $H_\mu(\xi\,|\,\eta)$ seen in equation (1.1) for the case of finite partitions is recovered. Henceforth we will not distinguish between a partition ξ and the σ-algebra $\sigma(\xi)$ that is generates.

4.1.2 Dependence on the Sub-σ-algebra Whose Information is Measured

In order to justify the definition of $H_\mu(\mathscr{C}\,|\,\mathscr{A})$ we need to know that $I_\mu(\mathscr{C}\,|\,\mathscr{A})$ is a measurable function. In addition to this, we shall see that both the information and the entropy are monotone and continuous (in a suitable sense) with respect to the σ-algebra whose information is being computed.

Proposition 4.5. *Let $(X, \mathscr{B}, \mu)$ be a Borel probability space with sub-σ-algebras $\mathscr{A}, \mathscr{C}, \mathscr{C}_n \subseteq \mathscr{B}$ for $n \in \mathbb{N}$, and assume that $\mathscr{C}, \mathscr{C}_n$ are countably generated for all $n \geqslant 1$. Then*

(1) *the map $x \mapsto I_\mu(\mathscr{C}\big|\mathscr{A})(x)$ is measurable;*
(2) *if $\mathscr{C}_1 \subseteq \mathscr{C}_2$ then $I_\mu(\mathscr{C}_1\big|\mathscr{A}) \leqslant I_\mu(\mathscr{C}_2\big|\mathscr{A})$; and*
(3) *if $\mathscr{C}_n \nearrow \mathscr{C}$ is an increasing sequence of σ-algebras then*

$$I_\mu(\mathscr{C}_n\big|\mathscr{A}) \nearrow I_\mu(\mathscr{C}\big|\mathscr{A})$$

and

$$H_\mu(\mathscr{C}_n\big|\mathscr{A}) \nearrow H_\mu(\mathscr{C}\big|\mathscr{A}).$$

Proof. Property (2) follows from the definition and the fact that $[x]_{\mathscr{C}_2} \subseteq [x]_{\mathscr{C}_1}$ for $x \in X$. For (1), first consider the case where $\mathscr{C} = \sigma(\xi)$ is generated by a countable partition $\xi = \{P_1, P_2, \dots\}$. In this case

$$I_\mu\left(\mathscr{C}\big|\mathscr{A}\right)(x) = \begin{cases} -\log \mu_x^{\mathscr{A}}(P_1) \text{ for } x \in P_1 \in \xi, \\ -\log \mu_x^{\mathscr{A}}(P_2) \text{ for } x \in P_2 \in \xi, \\ \quad\vdots \end{cases}$$

so $I_\mu(\mathscr{C}\big|\mathscr{A})$ is measurable. The general case of a countably-generated σ-algebra $\mathscr{C} = \sigma(\{C_1, C_2, \dots\})$ follows by defining the sequence (ξ_n) of partitions to have the property that $\mathscr{C}_n = \sigma(\xi_n) = \sigma(\{C_1, \dots, C_n\}) \nearrow \mathscr{C}$ and then applying (3), which we now prove.

Let $\mathscr{C}_n \nearrow \mathscr{C}$ be an increasing sequence of countably-generated σ-algebras. Then

$$[x]_{\mathscr{C}} = \bigcap_{n \geqslant 1} [x]_{\mathscr{C}_n}$$

and so

$$\mu_x^{\mathscr{A}}\left([x]_{\mathscr{C}_n}\right) \searrow \mu_x^{\mathscr{A}}\left([x]_{\mathscr{C}}\right)$$

which gives (3) by the definition of $I_\mu(\mathscr{C}\big|\mathscr{A})$. □

The next lemma gives an alternative description of conditional entropy by showing that $H_\mu(\mathscr{A}\big|\mathscr{C})$ is the average of the entropies $H_{\mu_x^{\mathscr{A}}}(\mathscr{C})$ (cf. equation (1.1)).

Lemma 4.6. *Let $(X, \mathscr{B}, \mu)$ be a Borel probability space, with $\mathscr{C}$ and $\mathscr{A}$ sub σ-algebras of $\mathscr{B}$. Then*

$$H_\mu\left(\mathscr{C}\big|\mathscr{A}\right) = \int H_{\mu_x^{\mathscr{A}}}(\mathscr{C})\, \mathrm{d}\mu(x),$$

where $H_{\mu_x^{\mathscr{A}}}(\mathscr{C})$ is infinite unless $\mathscr{C}$ agrees modulo $\mu_x^{\mathscr{A}}$ with a σ-algebra generated by a countable partition of finite entropy with respect to $\mu_x^{\mathscr{A}}$.

PROOF. By Proposition 4.5 and monotone convergence, it is enough to check the first statement for $\mathscr{C} = \sigma(\xi)$ for a finite partition ξ. In this case the properties of conditional expectation and conditional measures show that

$$\mu_x^{\mathscr{A}}(P)\log\mu_x^{\mathscr{A}}(P) = E(\chi_P\big|\mathscr{A})(x)\log\mu_x^{\mathscr{A}}(P) = E(\chi_P\log\mu_x^{\mathscr{A}}(P)\big|\mathscr{A})(x)$$

for any $P \in \xi$ and almost every $x \in X$. Therefore,

$$\begin{aligned}\int H_{\mu_x^{\mathscr{A}}}(\xi)\,\mathrm{d}\mu &= -\int \sum_{P\in\xi} \mu_x^{\mathscr{A}}(P)\log\mu_x^{\mathscr{A}}(P)\,\mathrm{d}\mu \\ &= -\int \sum_{P\in\xi} \chi_P \log\mu_x^{\mathscr{A}}(P)\,\mathrm{d}\mu \\ &= \int I_\mu\left(\xi\big|\mathscr{A}\right)\,\mathrm{d}\mu = H_\mu\left(\xi\big|\mathscr{A}\right).\end{aligned}$$

Finiteness of $H_{\mu_x^{\mathscr{A}}}(\mathscr{C})$ for a general σ-algebra $\mathscr{C}$ was discussed in Example 4.4. □

The next proposition gives a natural interpretation of zero conditional entropy.

Proposition 4.7. *Let $(X, \mathscr{B}, \mu)$ be a probability space, with $\mathscr{C}$ and $\mathscr{A}$ a pair of countably-generated sub σ-algebras of $\mathscr{B}$. Then*

$$H_\mu\left(\mathscr{C}\big|\mathscr{A}\right) = 0$$

if and only if

$$\mathscr{C} \underset{\mu}{\subseteq} \mathscr{A},$$

which means in this setting[25] *that there exists a μ-null set $N \subseteq X$ such that $C \in \mathscr{C}$ implies that $C \setminus N \in \mathscr{A}$.*

PROOF. Clearly $H_\mu\left(\mathscr{C}\big|\mathscr{A}\right) = 0 \iff \mu_x^{\mathscr{A}}\left([x]_{\mathscr{C}}\right) = 1$ for almost every $x \in X$. Now, using the fact that both σ-algebras are countably-generated, $\mathscr{C} \underset{\mu}{\subseteq} \mathscr{A}$ implies that there is a null set N with $[x]_{\mathscr{A}} \setminus N \subseteq [x]_{\mathscr{C}}$, so

$$\mu_x^{\mathscr{A}}\left([x]_{\mathscr{C}}\right) \geqslant \mu_x^{\mathscr{A}}\left([x]_{\mathscr{A}} \setminus N\right) = 1 \text{ a.e.}$$

If $\mu_x^{\mathscr{A}}\left([x]_{\mathscr{C}}\right) = 1$ almost everywhere, then for $C \in \mathscr{C}$

$$\mu_x^{\mathscr{A}}(C) = \chi_C(x)$$

almost everywhere, so

$$E\left(\chi_C\big|\mathscr{A}\right) = \chi_C.$$

This shows that $C \underset{\mu}{\in} \mathscr{A}$; that is, there exists $A \in \mathscr{A}$ with $\mu(A \triangle C) = 0$. Using this for the countably many generators of $\mathscr{C}$, and collecting the countably many null sets, we obtain $\mathscr{C} \underset{\mu}{\subseteq} \mathscr{A}$. □

4.1.3 Dependence on the Given Sub-σ-algebra

We now turn to continuity properties of information and entropy with respect to the given σ-algebra — that is, properties of the function

$$\mathscr{A} \mapsto I_\mu(\mathscr{C}\big|\mathscr{A})$$

for fixed $\mathscr{C}$. This is considerably more delicate than the corresponding properties for the function

$$\mathscr{C} \mapsto I_\mu(\mathscr{C}\big|\mathscr{A})$$

for fixed $\mathscr{A}$, considered in Section 4.1.2. In particular, we will need to assume that $\mathscr{C} = \sigma(\xi)$ for some partition ξ of finite entropy with respect to μ.

Proposition 4.8. *Let ξ be a countable partition of the Borel probability space $(X, \mathscr{B}, \mu)$ with $H_\mu(\xi) < \infty$. Let $\mathscr{A}_n \nearrow \mathscr{A}_\infty$ be an increasing sequence of σ-algebras. Then*

$$\int \sup_{n\geqslant 1} I_\mu\left(\xi\big|\mathscr{A}_n\right)\,\mathrm{d}\mu < \infty, \tag{4.3}$$

$$I_\mu\left(\xi\big|\mathscr{A}_n\right) \to I_\mu\left(\xi\big|\mathscr{A}_\infty\right) \tag{4.4}$$

almost everywhere and in L^1_μ, and

$$H_\mu\left(\xi\big|\mathscr{A}_n\right) \searrow H_\mu\left(\xi\big|\mathscr{A}_\infty\right) \tag{4.5}$$

as $n \to \infty$.

Proof. By the increasing martingale theorem (see [38, Th. 5.5]), for any atom P in ξ,

$$I_\mu\left(\xi\big|\mathscr{A}_n\right)(x) = -\log E_\mu\left(\chi_P\big|\mathscr{A}_n\right)(x) \to I_\mu\left(\xi\big|\mathscr{A}_\infty\right)(x)$$

for almost every $x \in P$. Thus, by the dominated convergence theorem, equation (4.4) and the convergence in equation (4.5) follow from equation (4.3). The monotonicity of the convergence will be shown in Proposition 4.9 below. Let

$$f = \sup_{n\geqslant 1} I_\mu\left(\xi\big|\mathscr{A}_n\right),$$

so in particular $f \geqslant 0$. Write $\mathbb{R}^+$ for the non-negative reals and m for Lebesgue measure restricted to $\mathbb{R}^+$. Let

$$F(t) = \mu\left(\{x \in X \mid f(x) > t\}\right);$$

then

$$\begin{aligned}\int f\,\mathrm{d}\mu &= \int_{X\times\mathbb{R}^+} \chi_{\{(x,t)|f(x)>t\}}\,\mathrm{d}(\mu\times m)\\ &= \int_0^\infty F(t)\,\mathrm{d}t\end{aligned}$$

by applying Fubini's theorem twice. Now

$$\begin{aligned} F(t) &= \mu\left(\left\{ x \mid \sup_{n\geqslant 1}\big(-\sum_{P\in\xi}\chi_P(x)\log\mu_x^{\mathscr{A}_n}(P)\ \big) > t\ \right\}\right) \\ &= \sum_{P\in\xi}\mu\left(\{x\in P \mid \inf_{n\geqslant 1}\mu_x^{\mathscr{A}_n}(P) < \mathrm{e}^{-t}\}\right) \\ &= \sum_{P\in\xi}\sum_{n=1}^{\infty}\mu\left(\{x\in P \mid \mu_x^{\mathscr{A}_n}(P) < \mathrm{e}^{-t} \text{ but } \mu_x^{\mathscr{A}_m}(P)\geqslant \mathrm{e}^{-t} \text{ for } m<n\}\right). \end{aligned}$$

For a fixed $t\geqslant 0$ and $P\in\xi$, let

$$A_n = \{x \mid \mu_x^{\mathscr{A}_n}(P) < \mathrm{e}^{-t} \text{ but } \mu_x^{\mathscr{A}_m}(P)\geqslant \mathrm{e}^{-t} \text{ for } m<n\}\in\mathscr{A}_n,$$

and note that

$$\mu(P\cap A_n) = \int_{A_n}\chi_P\,\mathrm{d}\mu < \mathrm{e}^{-t}\mu(A_n).$$

Then (notice that each A_n depends on P)

$$F(t)\leqslant \sum_{P\in\xi}\sum_{n=1}^{\infty}\mu\left(P\cap A_n\right)\leqslant \sum_{P\in\xi}\min\{\mu(P),\mathrm{e}^{-t}\},$$

which by integrating yields

$$\begin{aligned} \int_0^{\infty}F(t)\,\mathrm{d}t &\leqslant \sum_{P\in\xi}\int_0^{\infty}\min\{\mu(P),\mathrm{e}^{-t}\}\,\mathrm{d}t \\ &= \sum_{P\in\xi} -\mu(P)\log\mu(P)+\mu(P) \\ &= H_\mu(\xi)+1. \end{aligned}$$

□

We now turn to generalizing Proposition 1.7 to this setting.

Proposition 4.9. *Let $(X,\mathscr{B},\mu)$ be a Borel probability space, and let $\mathscr{A},\mathscr{C}_1,\mathscr{C}_2$ be countably-generated sub σ-algebras of $\mathscr{B}$. Then*

(1) $H_\mu\left(\mathscr{C}_1\vee\mathscr{C}_2\middle|\mathscr{A}\right) = H_\mu\left(\mathscr{C}_1\middle|\mathscr{A}\right) + H_\mu\left(\mathscr{C}_2\middle|\mathscr{A}\vee\mathscr{C}_1\right)$, *and similarly for the information function* $I_\mu(\cdot|\cdot)$;
(2) $H_\mu\left(\mathscr{C}_2\middle|\mathscr{A}\vee\mathscr{C}_1\right)\leqslant H_\mu\left(\mathscr{C}_2\middle|\mathscr{A}\right)$ *and, if* $H_\mu(\mathscr{C}_2)<\infty$, *then*

$$H_\mu(\mathscr{C}_2|\mathscr{C}_1) = H_\mu(\mathscr{C}_2)$$

if and only if $\mathscr{C}_1\perp\mathscr{C}_2$;
(3) $H_\mu\left(\mathscr{C}_1\vee\mathscr{C}_2\middle|\mathscr{A}\right)\leqslant H_\mu\left(\mathscr{C}_1\middle|\mathscr{A}\right) + H_\mu\left(\mathscr{C}_2\middle|\mathscr{A}\right)$.

Before continuing with the proof of Proposition 4.9 we record a connection between the behavior of conditional measures and independence.

Lemma 4.10. *Let $\mathscr{C}$ be a countably-generated σ-algebra in a Borel probability space $(X, \mathscr{B}, \mu)$, and let $P \in \mathscr{B}$ be any measurable set. Then P is independent of $\mathscr{C}$, meaning*

$$\mu(P \cap C) = \mu(P)\mu(C) \tag{4.6}$$

for all $C \in \mathscr{C}$, if and only if

$$\mu_x^{\mathscr{C}}(P) = \mu(P) \tag{4.7}$$

for almost every $x \in X$.

Proof. Suppose that equation (4.7) holds, and let $C \in \mathscr{C}$. Then

$$\mu(P \cap C) = \int_C \chi_P \, \mathrm{d}\mu = \int_C \underbrace{E_\mu\left(\chi_P \middle| \mathscr{C}\right)}_{=\mu_x^{\mathscr{C}}(P)} \mathrm{d}\mu = \mu(P)\mu(C).$$

Now assume equation (4.6). Then the constant function $f(x) = \mu(P)$ satisfies

$$\int_C f \, \mathrm{d}\mu = \mu(P)\mu(C) = \mu(P \cap C) = \int_C \chi_P \, \mathrm{d}\mu$$

for all $C \in \mathscr{C}$ and is clearly $\mathscr{C}$-measurable. It follows that

$$f(x) = E\left(\chi_P \middle| \mathscr{C}\right) = \mu_x^{\mathscr{C}}(P) \quad \mu\text{-a.e.}$$

□

Proof of Proposition 4.9. For (1), pick sequences of finite partitions (ξ_ℓ), (η_m) and (η'_n) with $\sigma(\xi_\ell) \nearrow \mathscr{C}_1$, $\sigma(\eta_m) \nearrow \mathscr{C}_2$, and $\sigma(\eta'_n) \nearrow \mathscr{A}$. First apply Proposition 4.8 and take $n \to \infty$ to deduce that

$$I_\mu(\xi_\ell \vee \eta_m \big| \mathscr{A}) = I_\mu(\xi_\ell \big| \sigma(\eta_m) \vee \mathscr{A}) + I_\mu(\eta_m \big| \mathscr{A}),$$

and then use Proposition 4.5 and 4.8 to take $m \to \infty$ and see that

$$I_\mu(\xi_\ell \vee \mathscr{C}_2 \big| \mathscr{A}) = I_\mu(\xi_\ell \big| \mathscr{C}_2 \vee \mathscr{A}) + I_\mu(\mathscr{C}_2 \big| \mathscr{A}),$$

and finally use Proposition 4.5 and take $\ell \to \infty$ to see that

$$I_\mu(\mathscr{C}_1 \vee \mathscr{C}_2 \big| \mathscr{A}) = I_\mu(\mathscr{C}_1 \big| \mathscr{C}_2 \vee \mathscr{A}) + I_\mu(\mathscr{C}_2 \big| \mathscr{A}),$$

proving (1).

For property (2) we start again with the case of a countable partition ξ with finite entropy in place of $\mathscr{C}_2$. Assume first that $\mathscr{A}$ is the trivial σ-algebra $\{\varnothing, X\}$. Then, using Lemma 4.6,

$$\begin{aligned} H_\mu\left(\xi\middle|\mathscr{C}\right) &= \int H_{\mu_x^{\mathscr{C}}}(\xi)\,\mathrm{d}\mu \\ &= -\sum_{P\in\xi}\int \phi\left(\mu_x^{\mathscr{C}}(P)\right)\,\mathrm{d}\mu \\ &\leqslant -\sum_{P\in\xi}\phi\Bigg(\underbrace{\int \mu_x^{\mathscr{C}}(P)\,\mathrm{d}\mu}_{=\mu(P)}\Bigg) \quad \text{(by Lemmas 1.3 and 1.4)} \\ &= H_\mu(\xi). \end{aligned}$$

By strict convexity of ϕ, equality occurs if and only if

$$\mu(P) = \mu_x^{\mathscr{C}}(P)$$

for almost every x and all $P \in \xi$, and this holds if and only if

$$\mu(P \cap C) = \mu(P)\mu(C)$$

for all $P \in \xi, C \in \mathscr{C}$ by Lemma 4.10. The general case will rely on the case above. Indeed, by Lemma 4.6, we have

$$\int H_{\mu_x^{\mathscr{A}}}\left(\mathscr{C}_2\middle|\mathscr{C}_1\right)\,\mathrm{d}\mu \leqslant \int H_{\mu_x^{\mathscr{A}}}\left(\mathscr{C}_2\right)\,\mathrm{d}\mu = H_\mu\left(\mathscr{C}_2\middle|\mathscr{A}\right).$$

We want to show that

$$\int H_{\mu_x^{\mathscr{A}}}\left(\mathscr{C}_2\middle|\mathscr{C}_1\right)\,\mathrm{d}\mu = H_\mu\left(\mathscr{C}_2\middle|\mathscr{C}_1 \vee \mathscr{A}\right),$$

which will prove (2). Expressing $H_{\mu_x^{\mathscr{A}}}(\mathscr{C}_2|\mathscr{C}_1)$ as an integral,

$$H_{\mu_x^{\mathscr{A}}}(\mathscr{C}_2\big|\mathscr{C}_1) = \int H_{(\mu_x^{\mathscr{A}})_y^{\mathscr{C}_1}}(\mathscr{C}_2)\,\mathrm{d}\mu_x^{\mathscr{A}}(y), \tag{4.8}$$

as in Lemma 4.6, we see that the proof is completed by noting that conditional measures "commute" with taking refinements in the sense that

$$(\mu_x^{\mathscr{A}})_y^{\mathscr{C}_1} = \mu_y^{\mathscr{A}\vee\mathscr{C}_1} \tag{4.9}$$

for almost every x and $\mu_x^{\mathscr{A}}$-almost every y, which holds by [38, Prop. 5.20]. This gives the result since

$$\begin{aligned} \int H_{\mu_y^{\mathscr{A}\vee\mathscr{C}_1}}(\mathscr{C}_2)\,\mathrm{d}\mu(y) &= \iint H_{\mu_y^{\mathscr{A}\vee\mathscr{C}_1}}(\mathscr{C}_2)\,\mathrm{d}\mu_x^{\mathscr{A}}(y)\,\mathrm{d}\mu(x) \\ &\qquad \text{(by [38, Th. 5.14])} \\ &= \iint H_{(\mu_x^{\mathscr{A}})_y^{\mathscr{C}_1}}(\mathscr{C}_2)\,\mathrm{d}\mu_x^{\mathscr{A}}(y)\,\mathrm{d}\mu(x) \\ &\qquad \text{(by equation (4.9))} \\ &= \int H_{\mu_x^{\mathscr{A}}}(\mathscr{C}_2\big|\mathscr{C}_1)\,\mathrm{d}\mu(x) \qquad \text{(by equation (4.8)).} \end{aligned}$$

Property (2) (and with it (3)) now follows from Proposition 4.5 once more. □

4.2 Conditional Entropy of a Transformation

We start with a direct generalization of Lemma 1.12.

Lemma 4.11. *Let $(X, \mathscr{B}, \mu, T)$ be a measure-preserving system. Then*

$$H_\mu\left(\mathscr{C}\middle|\mathscr{A}\right) = H_\mu\left(T^{-1}\mathscr{C}\middle|T^{-1}\mathscr{A}\right)$$

and

$$I_\mu\left(\mathscr{C}\middle|\mathscr{A}\right)\circ T = I_\mu\left(T^{-1}\mathscr{C}\middle|T^{-1}\mathscr{A}\right). \tag{4.10}$$

Proof. It is enough to show equation (4.10), and for this we have

$$\begin{aligned} I_\mu\left(\mathscr{C}\middle|\mathscr{A}\right)(Tx) &= -\log \mu_{Tx}^{\mathscr{A}}\left([Tx]_{\mathscr{C}}\right) \\ &= -\log\left(T_*\mu_x^{T^{-1}\mathscr{A}}\right)\left([Tx]_{\mathscr{C}}\right) \quad \text{(by [38, Cor. 5.24])} \\ &= -\log \mu_x^{T^{-1}\mathscr{A}}\left(T^{-1}[Tx]_{\mathscr{C}}\right) \\ &= -\log \mu_x^{T^{-1}\mathscr{A}}\left([x]_{T^{-1}\mathscr{C}}\right) \\ &= I_\mu\left(T^{-1}\mathscr{C}\middle|T^{-1}\mathscr{A}\right)(x). \end{aligned}$$

□

For later developments, it will be useful to discuss the entropy of T with respect to a given sub-σ-algebra $\mathscr{A}$ with $T^{-1}\mathscr{A} \subseteq \mathscr{A}$, so when measuring the entropies of the repeated experiment ξ we will always assume that the information of $\mathscr{A}$ is given. The first step is to show that the sequence of (conditional) entropies of the repeated experiments is subadditive. Given $m, n \geqslant 1$, and using Proposition 1.7(3) and Lemma 4.11 again,

$$\begin{aligned} H_\mu\left(\bigvee_{i=0}^{m+n-1} T^{-i}\xi\middle|\mathscr{A}\right) &= H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\middle|\mathscr{A}\right) \\ &\qquad + H_\mu\left(\bigvee_{i=n}^{m+n-1} T^{-i}\xi\middle|\mathscr{A}\vee\bigvee_{i=0}^{n-1} T^{-1}\xi\right) \\ &\leqslant H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\middle|\mathscr{A}\right) + H_\mu\left(T^{-n}\bigvee_{i=0}^{m-1} T^{-i}\xi\middle|T^{-n}\mathscr{A}\right) \\ &= H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\middle|\mathscr{A}\right) + H_\mu\left(\bigvee_{i=0}^{m-1} T^{-i}\xi\middle|\mathscr{A}\right), \end{aligned}$$

so the sequence $\left(H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\middle|\mathscr{A}\right)\right)_{n\geqslant 1}$ is subadditive in the sense of Lemma 1.13, justifying the claimed convergence in Definition 4.12. Here (and below) we write $\eta\vee\mathscr{A}$ as an abbreviation for $\sigma(\eta)\vee\mathscr{A}$.

Definition 4.12. *Let $(X, \mathscr{B}, \mu, T)$ be a measure-preserving system and let ξ be a partition of X with finite entropy. Then the entropy of T is*

$$h_\mu(T) = \sup_{\xi : H_\mu(\xi) < \infty} h_\mu(T, \xi).$$

If $\mathscr{A}$ is a sub σ-algebra of $\mathscr{B}$ with $T^{-1}(\mathscr{A}) \subseteq \mathscr{A}$ then the conditional entropy of T given $\mathscr{A}$ is

$$h_\mu(T\big|\mathscr{A}) = \sup_{\xi : H_\mu(\xi) < \infty} h_\mu(T, \xi\big|\mathscr{A})$$

where

$$h_\mu(T, \xi\big|\mathscr{A}) = \lim_{n\to\infty} \frac{1}{n} H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}(\xi)\big|\mathscr{A}\right) = \inf_{n\geqslant 1} \frac{1}{n} H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}(\xi)\big|\mathscr{A}\right).$$

Just as in Chapter 1, we need to develop the basic properties of conditional entropy.

Proposition 4.13. *Let $(X, \mathscr{B}, \mu, T)$ be a measure-preserving system on a Borel probability space, and let ξ and η be countable partitions of X with finite entropy. Then properties* (1) *to* (5) *of Proposition* 1.16 *also hold for conditional entropies conditioned on $\mathscr{A}$. Moreover,*

(1) $h_\mu(T, \xi\big|\mathscr{A}) = \lim_{n\to\infty} H_\mu\left(\xi\big|\bigvee_{i=1}^{n} T^{-i}\xi \vee \mathscr{A}\right) = H_\mu\left(\xi\big|\bigvee_{i=1}^{\infty} T^{-i}\xi \vee \mathscr{A}\right)$;
(2) *if T is invertible, then*

$$\begin{aligned} h_\mu\left(T, \xi \vee \eta\big|\mathscr{A}\right) &= h_\mu(T, \xi\big|\mathscr{A}) + h_\mu\left(T, \eta\big|\bigvee_{-\infty}^{\infty} T^{-i}\xi \vee \mathscr{A}\right) \\ &= h_\mu(T, \xi\big|\mathscr{A}) + H_\mu\left(\eta\big|\bigvee_{i=1}^{\infty} T^{-i}\eta \vee \bigvee_{-\infty}^{\infty} T^{-i}\xi \vee \mathscr{A}\right); \end{aligned}$$

Proof. The proof of the generalization of Proposition 1.16 to conditional entropies follows the same lines as that proof, giving properties (1) to (5) of Proposition 1.16 easily.
(1): For any $n \geqslant 1$,

$$\begin{aligned}
\frac{1}{n}H_\mu\left(\bigvee_{i=0}^{n-1}T^{-i}\xi\vee\mathscr{A}\right) &= \frac{1}{n}\left(H_\mu\left(\xi\Big|\bigvee_{i=1}^{n-1}T^{-i}\xi\vee\mathscr{A}\right)\right.\\
&\qquad\left.+H_\mu\left(T^{-1}\bigvee_{i=0}^{n-2}T^{-i}\xi\vee\mathscr{A}\right)\right)\\
&= \frac{1}{n}\left(H_\mu\left(\xi\Big|\bigvee_{i=1}^{n-1}T^{-i}\xi\vee\mathscr{A}\right)\right.\\
&\qquad+H_\mu\left(T^{-1}\xi\Big|\bigvee_{i=2}^{n-1}T^{-i}\xi\vee\mathscr{A}\right)\\
&\qquad\left.+\cdots+H_\mu\left(T^{-(n-1)}\xi\Big|\vee\mathscr{A}\right)\right)\\
&= \frac{1}{n}\sum_{j=0}^{n-1}H_\mu\left(\xi\Big|\bigvee_{i=1}^{j}T^{-i}\xi\vee\mathscr{A}\right)\\
&\underset{\text{Cesàro}}{\longrightarrow} H_\mu\left(\xi\Big|\bigvee_{i=1}^{\infty}T^{-i}\xi\vee\mathscr{A}\right).
\end{aligned}$$

In the last line we first used Lemma 4.11, then Proposition 4.8 and the fact that the Cesàro averages $\frac{1}{n}\sum_{j=0}^{n-1}a_j$ of a convergent sequence (a_m) converge to the limit $\lim_{m\to\infty}a_m$ as $n\to\infty$.
(2): By splitting the entropy up in a similar way to the argument in (1), but in a different order, we get

$$\begin{aligned}
h_\mu(T,\xi\vee\eta\big|\mathscr{A}) &= \lim_{n\to\infty}\frac{1}{n}H_\mu\left(\bigvee_{i=0}^{n-1}T^{-i}\xi\vee\mathscr{A}\right)\\
&\qquad+\lim_{n\to\infty}\frac{1}{n}H_\mu\left(\bigvee_{j=0}^{n-1}T^{-j}\eta\Big|\bigvee_{i=0}^{n-1}T^{-i}\xi\vee\mathscr{A}\right)\\
&= h_\mu(T,\xi\big|\mathscr{A})\\
&\qquad+\lim_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n-1}H_\mu\left(T^{-j}\eta\Big|\bigvee_{i=j+1}^{n-1}T^{-i}\eta\vee\bigvee_{i=0}^{n-1}T^{-i}\xi\vee\mathscr{A}\right)\\
&= h_\mu(T,\xi\big|\mathscr{A})\\
&\qquad+\lim_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n-1}H_\mu\left(\eta\Big|\bigvee_{i=1}^{n-1-j}T^{-i}\eta\vee\bigvee_{i=-j}^{n-1-j}T^{-i}\xi\vee\mathscr{A}\right)\\
&\geqslant h_\mu(T,\xi\big|\mathscr{A})+H_\mu\left(\eta\Big|\bigvee_{i=1}^{\infty}T^{-i}\eta\vee\bigvee_{i=-\infty}^{\infty}T^{-i}\xi\vee\mathscr{A}\right).
\end{aligned}$$

On the other hand, given $\varepsilon > 0$ there exists N such that

$$H_\mu\left(\eta\Big|\bigvee_{i=1}^{N} T^{-i}\eta \vee \bigvee_{i=-N}^{N} T^{-i} \vee \mathscr{A}\xi\right) < H_\mu\left(\eta\Big|\bigvee_{i=1}^{\infty} T^{-i}\eta \vee \bigvee_{i=-\infty}^{\infty} T^{-i}\xi \vee \mathscr{A}\right) + \varepsilon,$$

by Proposition 4.8. Therefore, for any n and $j \in [N, n-N-1]$,

$$H_\mu\left(\eta\Big|\bigvee_{i=1}^{n-1-j} T^{-i}\eta \vee \bigvee_{i=-j}^{n-1-j} T^{-i}\xi \vee \mathscr{A}\right) \leqslant H_\mu\left(\eta\Big|\bigvee_{i=1}^{\infty} T^{-i}\eta \vee \bigvee_{i=-\infty}^{\infty} T^{-1}\xi \vee \mathscr{A}\right) + \varepsilon.$$

Since for large enough n the contribution of the other terms, which is at most $2NH_\mu(\eta)$, will eventually become smaller than $n\varepsilon$ as $n \to \infty$, the reverse inequality follows. □

Theorem 1.20 transfers some of the difficulty inherent in computing entropy onto the problem of finding a generator. There are general results[(26)] showing that generators always exist under suitable conditions (notice that the existence of a generator with k atoms means the entropy cannot exceed $\log k$), but these are of little direct help in constructing a generator. In Section 4.5 a generator will be found for a non-trivial example. The next result is a useful alternative to Theorem 1.20 in situations where a generator is difficult to find.

Theorem 4.14. *Suppose that $(X, \mathscr{B}, \mu, T)$ is a measure-preserving system on a Borel probability space. If (ξ_n) is an increasing sequence of partitions (this means that the atoms of ξ_n are members of $\sigma(\xi_{n+1})$ for all $n \geqslant 1$) of finite entropy with the property that*

- $\mathscr{B} \underset{\mu}{=} \bigvee_{i=1}^{\infty} \bigvee_{n=0}^{\infty} T^{-n}\sigma(\xi_i)$ *or*
- $\mathscr{B} \underset{\mu}{=} \bigvee_{i=1}^{\infty} \bigvee_{n=-\infty}^{\infty} T^{-n}\sigma(\xi_i)$ *if T is invertible,*

then

$$h_\mu(T) = \sup_i h_\mu(T, \xi_i) = \lim_{i\to\infty} h_\mu(T, \xi_i).$$

More generally, under the same hypothesis, if $\mathscr{A} \subseteq \mathscr{B}$ is a sub-σ-algebra with $T^{-1}\mathscr{A} \subseteq \mathscr{A}$, then

$$h_\mu(T\big|\mathscr{A}) = \sup_i h_\mu(T, \xi_i\big|\mathscr{A}) = \lim_{i\to\infty} h_\mu(T, \xi_i\big|\mathscr{A}).$$

The property of the sequence of partitions (ξ_n) is called *generating under T*.

Proof of Theorem 4.14. Let $\mathscr{A}$ and (ξ_i) be as in the theorem, and let ξ be another countable partition of finite entropy. Then

$$h_\mu(T, \xi\big|\mathscr{A}) \leqslant h_\mu(T, \bigvee_{j=0}^{n} T^{-j}\xi_i\big|\mathscr{A}) + H_\mu\left(\xi\Big|\bigvee_{j=0}^{n} T^{-j}\xi_i \vee \mathscr{A}\right)$$

by Proposition 1.16(3) conditioned on $\mathscr{A}$. Here the first term

$$h_\mu(T, \bigvee_{j=0}^{n} T^{-j}\xi_i \big| \mathscr{A}) = h_\mu(T, \xi_i \big| \mathscr{A})$$

by Proposition 1.16(4), and the second term

$$H_\mu\left(\xi \big| \bigvee_{j=0}^{n} T^{-j}\xi_i \vee \mathscr{A}\right) \to 0$$

as $i \to \infty$ by Proposition 4.8 and the assumption about the sequence (ξ_i). □

We end this section by showing how the entropy of a measure-preserving transformation splits into two terms when the transformation has a factor[(27)].

Corollary 4.15 (Abramov–Rokhlin formula). *Let $\mathsf{X} = (X, \mathscr{B}, \mu, T)$ be a measure-preserving system on a Borel probability space with T invertible. Then, for any factor (Y, ν, S) of X,*

$$h_\mu(T) = h_\nu(S) + h_\mu(T \big| \mathscr{A}) \tag{4.11}$$

where the factor Y is identified with its corresponding invariant sub-σ-algebra

$$\mathscr{A} = \underset{\mu}{T^{-1}\mathscr{A}} \subseteq \mathscr{B}.$$

Moreover,

$$h_\nu(S) = \sup\{h_\mu(T, \xi) \mid \xi \subseteq \mathscr{A} \text{ is a countable partition of finite entropy}\}.$$

Proof. Let $\phi : X \to Y$ be the factor map. For the final statement, let $\eta \subseteq \mathscr{A}$ be a countable partition. Then $H_\nu(\eta) = H_\mu(\phi^{-1}\eta)$ so that, in particular, η has finite entropy if and only if $\phi^{-1}\eta$ has finite entropy. Using this observation for $\bigvee_{i=0}^{n-1} S^{-1}\eta$ instead of η, we see that

$$h_\nu(S, \eta) = h_\mu(T, \phi^{-1}\eta),$$

which implies the last statement.

Turning to the proof of equation (4.11), pick sequences of finite partitions (η_m) and (ξ_n) with $\sigma(\eta_m) \nearrow \mathscr{A}$ and $\sigma(\xi_n) \nearrow \mathscr{B}$. By Proposition 4.13(2),

$$h_\mu(T, \xi_n \vee \phi^{-1}\eta_m) = h_\mu(T, \phi^{-1}\eta_m) + H_\mu\left(\xi_n \big| \bigvee_{i=1}^{\infty} T^{-i}\xi_n \vee \bigvee_{i=-\infty}^{\infty} T^{-i}\phi^{-1}\eta_m\right).$$

Notice that $h_\mu(T, \phi^{-1}\eta_m) = h_\nu(S, \eta_m)$. By Theorem 4.14, for any fixed $\varepsilon > 0$ there is some n such that

$$h_\mu(T) - \varepsilon \leqslant h_\mu(T, \xi_n) \leqslant h_\mu(T, \xi_n \vee \phi^{-1}\eta_m) \leqslant h_\mu(T)$$

and similarly

$$h_\mu(T\big|\mathscr{A}) - \varepsilon \leqslant h_\mu(T, \xi_n\big|\mathscr{A}).$$

Let $m \to \infty$ and apply Theorem 4.14 and Proposition 4.8 to obtain

$$h_\mu(T) - \varepsilon \leqslant h_\nu(S) + H_\mu\left(\xi_n\Big|\bigvee_{i=1}^{\infty} T^{-1}\xi_n \vee \mathscr{A}\right) \leqslant h_\mu(T).$$

Thus

$$H_\mu\left(\xi_n\Big|\bigvee_{i=1}^{\infty} T^{-i}\xi_n \vee \mathscr{A}\right) = h_\mu(T, \xi_n\big|\mathscr{A}) \leqslant h_\mu(T\big|\mathscr{A}),$$

so

$$\begin{aligned} h_\mu(T) - \varepsilon &\leqslant h_\nu(S) + h_\mu(T\big|\mathscr{A}) \\ &\leqslant h_\nu(S) + h_\mu(T, \xi_n\big|\mathscr{A}) + h_\mu(T\big|\mathscr{A}) - h_\mu(T, \xi_n\big|\mathscr{A}) \\ &\leqslant h_\mu(T) + \varepsilon \end{aligned}$$

as required. □

Exercises for Section 4.1

Exercise 4.2.1. Show that Proposition 4.8 does not hold for an arbitrary σ-algebra $\mathscr{C}$ and sequence of σ-algebras $\mathscr{A}_n \nearrow \mathscr{A}_\infty$, by finding an example for which $H_\mu\left(\mathscr{C}\big|\mathscr{A}_n\right) \not\to H_\mu\left(\mathscr{C}\big|\mathscr{A}_\infty\right)$.

Exercise 4.2.2. Prove Theorem 4.14.

4.3 The Pinsker Algebra

Decomposing a measure-preserving system into simpler constituents can have far-reaching consequences (a striking instance of this is Furstenberg's proof of Szemerédi's theorem; see [38, Chap. 7]). It is natural to ask if a measure-preserving system can always be decomposed into a zero-entropy system and a system with the property that every non-trivial factor has positive entropy. This turns out to be too much to ask, but a partial answer in the same direction is afforded by the Pinsker algebra [115].

Definition 4.16. *Let $(X, \mathscr{B}, \mu, T)$ be an invertible measure-preserving system on a Borel probability space. The* Pinsker algebra *of T is*

$$\begin{aligned} \mathscr{P}(T) &= \{B \in \mathscr{B} \mid h_\mu(T, \{B, X \setminus B\}) = 0\} \\ &= \left\{B \in \mathscr{B} \mid B \underset{\mu}{\subseteq} \bigvee_{i=1}^{\infty} \sigma\left(T^{-i}\{B, X \setminus B\}\right)\right\}. \end{aligned} \tag{4.12}$$

The formulation of $\mathscr{P}(T)$ in equation (4.12) may be described as follows: the Pinsker algebra comprises those sets with the property that knowledge of whether the orbit of a point lies in the set in all of the future determines whether it lies in the set in the present.

Proposition 4.17. *The Pinsker algebra $\mathscr{P}(T)$ is a T-invariant σ-algebra and so defines the* Pinsker factor *of T. This factor $\mathsf{X}_{\mathscr{P}} = (X_{\mathscr{P}}, \mathscr{P}(T), \mu, T_{\mathscr{P}})$ has zero entropy and is maximal with respect to that property: if $\mathsf{Y} = (Y, \mathscr{A}, \nu, S)$ is another factor of X with zero entropy then Y is a factor of $\mathsf{X}_{\mathscr{P}}$. Moreover, the relative entropy of T given $\mathscr{P}(T)$ coincides with the entropy of T,*

$$h_\mu(T) = h_\mu(T \big| \mathscr{P}(T)). \tag{4.13}$$

Proof. Recall the correspondence between invariant σ-algebras and factors discussed in [38, Sect. 6.2]. To see that $\mathscr{P}(T)$ is a σ-algebra, let $\{B_i \mid i \geqslant 1\}$ be a collection of sets in $\mathscr{P}(T)$ and write $\xi_i = \{B_i, X \setminus B_i\}$ for the associated partitions. If $Q \in \bigvee_{i=1}^{\infty} \sigma(\xi_i)$ and $\eta = \{Q, X \setminus Q\}$, then for any $\varepsilon > 0$ there is an n such that

$$H\left(\eta \Big| \bigvee_{i=1}^{n} \xi_i\right) < \varepsilon.$$

It follows, by Proposition 1.16(3), that

$$\begin{aligned} h(T, \eta) &\leqslant h(T, \bigvee_{i=1}^{n} \xi_i) + H(\eta \Big| \bigvee_{i=1}^{n} \xi_i) \\ &\leqslant \sum_{i=1}^{n} h(T, \xi_i) + \varepsilon = \varepsilon, \end{aligned}$$

so $\eta \subseteq \mathscr{P}(T)$. The T-invariance is clear, since in fact $h_\mu(T, \xi) = h_\mu(T, T^{-1}\xi)$ for any partition with finite entropy. By the definition of $\mathscr{P}(T)$ and by Corollary 4.15, the entropy $h_\mu(T_{\mathscr{P}})$ of the Pinsker factor vanishes. Now equation (4.13) follows from Corollary 4.15. If Y is any factor of X with zero entropy, then every finite partition of the corresponding T-invariant σ-algebra $\mathscr{A}$ must have zero entropy. Hence the whole σ-algebra $\mathscr{A}$ must be measurable with respect to $\mathscr{P}(T)$ modulo μ. By [38, Th. 6.5], it follows that there is a factor map $X_{\mathscr{P}} \to Y$. □

Theorem 4.18. *For any invertible measure-preserving transformation T of a Borel probability space $(X, \mathscr{B}, \mu)$,*

$$\mathscr{P}(T) = \bigvee_{\xi : H_\mu(\xi) < \infty} \bigcap_{n=0}^{\infty} \bigvee_{i=n}^{\infty} \sigma\left(T^{-i}(\xi)\right).$$

The σ-algebra $\bigcap_{n=0}^{\infty}\bigvee_{i=n}^{\infty}\sigma\left(T^{-i}(\xi)\right)$ is called the *tail σ-algebra* or *tail field* of the partition ξ.

Proof of Theorem 4.18. Let ξ be a partition with $H_\mu(\xi)<\infty$, and let η be a finite partition measurable with respect to $\bigcap_{n=0}^{\infty}\bigvee_{i=n}^{\infty}T^{-i}(\xi)$. Then

$$\begin{aligned}
h_\mu(T,\xi) &= H_\mu\left(\xi\Big|\bigvee_{i=1}^{\infty}T^{-i}\xi\right)\\
&= H_\mu\left(\xi\vee\eta\Big|\bigvee_{i=1}^{\infty}T^{-i}(\xi\vee\eta)\right)\\
&= h_\mu(T,\xi\vee\eta)\\
&= h_\mu(T,\eta)+h_\mu\left(T,\xi\Big|\bigvee_{i=-\infty}^{\infty}T^{-i}\eta\right)\\
&= h_\mu(T,\eta)+h_\mu(T,\xi),
\end{aligned}$$

where the last equality holds since

$$h_\mu\left(T,\xi\Big|\bigvee_{i=-\infty}^{\infty}T^{-i}\eta\right)=H_\mu\left(\xi\Big|\bigvee_{i=1}^{\infty}T^{-i}\xi\vee\bigvee_{i=-\infty}^{\infty}T^{-i}\eta\right)$$

and

$$\bigvee_{i=-\infty}^{\infty}T^{-i}\eta\leqslant\bigvee_{i=1}^{\infty}T^{-i}\xi.$$

It follows that $h_\mu(T,\eta)=0$ since $h_\mu(T,\xi)\leqslant H_\mu(\xi)<\infty$.

Conversely, if $\eta=\{Q,X\setminus Q\}\subseteq\mathscr{P}(T)$ then

$$h_\mu(T,\eta)=0=H_\mu(\eta|\bigvee_{i=1}^{\infty}T^{-i}\eta)$$

so $\eta\leqslant\bigvee_{i=1}^{\infty}T^{-i}\eta$ (modulo null sets). In particular,

$$\bigvee_{i=0}^{\infty}T^{-i}\eta=\bigvee_{i=1}^{\infty}T^{-i}\eta=\bigvee_{i=n}^{\infty}T^{-i}\eta$$

for all $n\geqslant 1$, which implies that

$$Q\in\bigcap_{n=0}^{\infty}\bigvee_{i=n}^{\infty}T^{-i}\eta$$

as required. □

If a generator is known, then the tail σ-algebra can be expressed in terms of the generator, giving the following strengthening of Theorem 4.18.

Theorem 4.19. *Let ξ be a generator for an invertible measure-preserving transformation T. Then*

$$\mathscr{P}(T) \underset{\mu}{=} \bigcap_{n=1}^{\infty} \bigvee_{k=n}^{\infty} T^{-k}(\xi),$$

the tail *of ξ. In particular, the σ-algebra of invariant sets $\mathscr{E}$ is a subset of $\bigvee_{k=1}^{\infty} T^{-k}(\xi)$.*

Example 4.20. Let $(X, \mathscr{B}, \mu, \sigma)$ be the Bernoulli shift defined by the probability vector $(p_1, \dots, p_s)$, so that $X = \prod_{\mathbb{Z}}\{1, \dots, s\}$, $\mu = \prod_{\mathbb{Z}}(p_1, \dots, p_s)$, and σ is the left shift. The state partition

$$\xi = \{[1]_0, [2]_0, \dots, [s]_0\}$$

is a generator, and the atoms of $\bigvee_{k=n}^{\infty} \sigma^{-k}(\xi)$ are sets of the form

$$A_n = \{x \in X \mid x_k = a_k \text{ for } k \geqslant n\};$$

it follows that the only non-empty set in the tail of ξ is X. Thus a Bernoulli shift has trivial Pinsker algebra[28].

Proof of Theorem 4.19. Write $\mathscr{A} = \bigvee_{i=0}^{\infty} T^{-i}\xi$. For any $n \geqslant 1$, the partition $\bigvee_{j=-n}^{-1} T^{-j}\xi$ is a generator for T^n, so by Proposition 1.16(5),

$$h_\mu(T^n) = H_\mu\left(\bigvee_{j=-n}^{-1} T^{-j}\xi \middle| \mathscr{A}\right)$$

and similarly

$$h_\mu(T^n \big| \mathscr{P}(T)) = H_\mu\left(\bigvee_{j=-n}^{-1} T^{-j}\xi \middle| \mathscr{A} \vee \mathscr{P}(T)\right).$$

On the other hand, by Proposition 4.17,

$$h_\mu(T^n) = nh_\mu(T) = nh_\mu(T \big| \mathscr{P}(T)) = h_\mu(T^n \big| \mathscr{P}(T)).$$

Thus for any partition $\eta \subseteq \mathscr{P}(T)$ we have, by Proposition 1.7,

$$\begin{aligned} H_\mu\left(\eta \middle| \mathscr{A}\right) &= H_\mu\left(\bigvee_{j=-n}^{-1} T^{-j}\xi \vee \eta \middle| \mathscr{A}\right) - H_\mu\left(\bigvee_{j=-n}^{-1} T^{-j}\xi \middle| \mathscr{A} \vee \sigma(\eta)\right) \\ &= \underbrace{H_\mu\left(\eta \middle| \mathscr{A} \vee \bigvee_{j=-n}^{-1} T^{-j}\xi\right)}_{<\varepsilon \text{ for large } n} \\ &\qquad + \underbrace{H_\mu\left(\bigvee_{j=-n}^{-1} T^{-j}\xi \middle| \mathscr{A}\right) - H_\mu\left(\bigvee_{j=-n}^{-1} T^{-j}\xi \middle| \mathscr{A} \vee \sigma(\eta)\right)}_{=0}. \end{aligned}$$

It follows that

$$H_\mu\left(\eta\middle|\mathscr{A}\right) = 0$$

for all $\eta \subseteq \mathscr{P}(T)$, so

$$H_\mu\left(\mathscr{P}(T)\middle|\bigvee_{i=0}^{\infty} T^{-i}\xi\right) = 0,$$

or

$$\mathscr{P}(T) \underset{\mu}{\subseteq} \bigvee_{i=0}^{\infty} T^{-i}\xi.$$

Since $\mathscr{P}(T)$ is T-invariant, we deduce that $\mathscr{P}(T)$ belongs to the tail of ξ (modulo null sets). □

A similar result holds for a sequence of σ-algebras that generate under the transformation in the limit.

Theorem 4.21. *If $(X, \mathscr{B}, \mu, T)$ is an invertible measure-preserving system on a Borel probability space, and $(\mathscr{P}_k)$ is a sequence of σ-algebras with*

$$\bigvee_{i=-\infty}^{\infty} T^{-i}\mathscr{P}_k \nearrow \mathscr{B},$$

then

$$\mathscr{P}(T) = \bigvee_{k\geqslant 1}\bigcap_{n=1}^{\infty}\bigvee_{i=n}^{\infty} T^{-i}\mathscr{P}_k.$$

Proof. Given any finite partition η and $\varepsilon > 0$, choose k so large that

$$H_\mu\left(\eta\middle|\bigvee_{i=-\infty}^{\infty} T^{-i}\mathscr{P}_k\right) < \varepsilon$$

and n so large that

$$H_\mu\left(\eta\middle|\bigvee_{i=-n}^{\infty} T^{-i}\mathscr{P}_k\right) < \varepsilon$$

and proceed as in the proof of Theorem 4.19. □

The next lemma (which generalizes equation (4.13)) encapsulates once again the idea that the factor $\mathscr{P}(T)$ has no entropy, and is maximal with respect to this property, so all the entropy of T must be visible relative to the factor $\mathscr{P}(T)$.

Lemma 4.22. *For an invertible measure-preserving transformation T on a Borel probability space and a partition ξ with finite entropy,*

$$h_\mu(T, \xi) = h_\mu(T, \xi\big|\,\mathscr{P}(T)).$$

PROOF. Let $\eta \subseteq \mathscr{P}(T)$ be a partition with $H_\mu(\eta) < \infty$ (so $h_\mu(T, \eta) = 0$) and let ξ be a partition with $H_\mu(\xi) < \infty$. Then

$$h_\mu(T, \xi) \leqslant h_\mu(T, \xi \vee \eta) = h_\mu(T, \eta) + h_\mu\left(T, \xi \middle| \bigvee_{j=-\infty}^{\infty} T^{-j}\eta\right) \leqslant h_\mu(T, \xi)$$

by Proposition 4.13(2), which with the above also implies that

$$h_\mu(T, \xi) = H_\mu\left(\xi \middle| \bigvee_{i=1}^{\infty} T^{-i}\xi \vee \bigvee_{i=-\infty}^{\infty} T^{-i}\eta\right).$$

By choosing an increasing sequence of such partitions (η_n), which generate the Pinsker algebra in the sense that $\sigma(\eta_n) \nearrow \mathscr{P}(T)$, the lemma follows from the continuity of entropy with respect to the given σ-algebra (see Proposition 4.8). □

4.4 Entropy and Convex Combinations

It is often useful to assume an invariant measure is ergodic – [38, Th. 6.2] shows how any invariant measure can be decomposed into ergodic components. In this section we show how entropy behaves with respect to generalized convex combinations, including the ergodic decomposition as a special case.

Theorem 4.23. *Let $(X, \mathscr{B}_X, \mu, T)$ be an invertible measure-preserving system on a Borel probability space, with ergodic decomposition*

$$\mu = \int_Y \mu_y \, \mathrm{d}\nu(y), \tag{4.14}$$

where $(Y, \mathscr{B}_Y, \nu)$ is some Borel probability space. Then

$$h_\mu(T) = \int_Y h_{\mu_y}(T) \, \mathrm{d}\nu(y) \tag{4.15}$$

and

$$h_\mu(T, \xi) = \int_Y h_{\mu_y}(T, \xi) \, \mathrm{d}\nu(y) \tag{4.16}$$

for any partition ξ with $H_\mu(\xi) < \infty$. The conclusion in equation (4.15) *also holds if equation* (4.14) *is any way of expressing μ as a generalized convex combination of invariant measures.*

We will give two related but slightly different proofs, the first for the ergodic decomposition exploits the construction of the ergodic decomposition using conditional measures with respect to the σ-algebra of invariant sets.

The second proof uses the Abramov–Rokhlin formula (Corollary 4.15) to deal with any generalized convex combination of measures.

Proof of equations (4.15) and (4.16) using the Pinsker algebra. Recall from [38, Sect. 6.1] that one way to construct the ergodic decomposition is to use $Y = X$ and $\mu_x = \mu_x^{\mathscr{E}}$ for $x \in X$, where

$$\mathscr{E} = \{E \in \mathscr{B}_X \mid T^{-1}E = E\}.$$

If $E \in \mathscr{E}$ then

$$T^{-1}\{E, X \setminus E\} = \{E, X \setminus E\}$$

so $E \in \mathscr{P}(T)$, the Pinsker algebra of T. If ξ is a partition with $H_\mu(\xi) < \infty$, then by Lemma 4.6, Lemma 4.22, and dominated convergence (alternatively using monotone convergence and Exercise 4.4.1) we have

$$\begin{aligned} h_\mu(T,\xi) &= h_\mu(T,\xi\big|\mathscr{E}) \\ &= \lim_{N\to\infty} \frac{1}{N} H_\mu\left(\bigvee_{n=0}^{N-1} T^{-n}\xi\big|\mathscr{E}\right) \\ &= \lim_{N\to\infty} \int \frac{1}{N} H_{\mu_x^{\mathscr{E}}}\left(\bigvee_{n=0}^{N-1} T^{-n}\xi\right) \mathrm{d}\mu(x) \\ &= \int h_{\mu_x^{\mathscr{E}}}(T,\xi)\,\mathrm{d}\mu(x). \end{aligned}$$

Now take a generating sequence of finite partitions (ξ_n) with

$$\sigma(\xi_n) \nearrow \mathscr{B}_X$$

to deduce, using Theorem 4.14 and monotone convergence of the entropy

$$h_{\mu_x^{\mathscr{A}}}(T,\xi_n)$$

in the integral above, that $h_\mu(T) = \int h_{\mu_x^{\mathscr{E}}}(T)\,\mathrm{d}\mu(x)$. □

Proof of equation (4.15) using the Abramov–Rokhlin formula. By assumption, we are given a Borel probability space $(Y, \mathscr{B}_Y, \nu)$ and a measurable function $y \mapsto \mu_y$, defined ν-almost everywhere, with μ_y a T-invariant Borel probability measure on $(X, \mathscr{B}_X)$, such that

$$\mu = \int_Y \mu_y \,\mathrm{d}\nu(y).$$

We define a probability measure ρ on $(Z, \mathscr{B}_Z) = (X \times Y, \mathscr{B}_X \otimes \mathscr{B}_Y)$ by

$$\rho = \int_Y \mu_y \times \delta_y \,\mathrm{d}\nu(y),$$

and note that if π_X, π_Y denote the projections onto the X and Y coordinates from Z, then $(\pi_X)_*\rho = \mu$ and $(\pi_Y)_*\rho = \nu$. We will use Corollary 4.15 to compute the entropy of the map $T \times I_Y : Z \to Z$ sending (x, y) to $(T(x), y)$, in two different ways. Firstly, $T \times I_Y$ is an extension of the identity map on Y, so

$$h_\rho(T \times I_Y) = h_\nu(I_Y) + h_\rho(T \times I_Y \big| \mathscr{N}_X \times \mathscr{B}_Y) \tag{4.17}$$

where $\mathscr{N}_X = \{\varnothing, X\}$ is the trivial σ-algebra on X; secondly, $T \times I_Y$ is an extension of the map T on $(X, \mathscr{B}_X, \mu)$, so

$$h_\rho(T \times I_Y) = h_\mu(T) + h_\rho(T \times I_Y \big| \mathscr{B}_X \times \mathscr{N}_Y). \tag{4.18}$$

Clearly $h_\nu(I_Y) = 0$ as in Example 1.23. We claim that

$$h_\rho(T \times I_Y \big| \mathscr{B}_X \times \mathscr{N}_Y) = 0$$

for the following reason. Let (ξ_n) be an increasing sequence of partitions of X with $\sigma(\xi_n) \nearrow \mathscr{B}_X$, and similarly let (η_n) be an increasing sequence of partitions of Y with $\sigma(\eta_n) \nearrow \mathscr{B}_Y$. Then

$$\begin{aligned} h_\rho(T \times I_Y, \xi_n \times \eta_n \big| \mathscr{B}_X \times \mathscr{N}_Y) = {} & h_\rho(T \times I_Y, \xi_n \times \{\varnothing, Y\} \big| \mathscr{B}_X \times \mathscr{N}_Y) \\ & + h_\rho(T \times I_Y, \{\varnothing, X\} \times \eta_n \big| \mathscr{B}_X \times \mathscr{N}_Y) \end{aligned}$$

by Corollary 4.15, and both terms vanish by Proposition 4.7. It follows from equation (4.17) and (4.18) that

$$h_\mu(T) = h_\rho(T \times I_Y \big| \mathscr{N}_X \times \mathscr{B}_Y).$$

Now

$$\rho^{\{\varnothing, X\} \times \mathscr{B}_Y}_{(x,y)} = \mu_y \times \delta_y \tag{4.19}$$

by the definition of ρ and [38, Prop. 5.19].

Let $\xi \subseteq \mathscr{B}_X$ and $\eta \subseteq \mathscr{B}_Y$ be finite partitions. Then

$$\begin{aligned} h_\rho(T \times I_Y, \xi \times \eta \big| \mathscr{N}_X \times \mathscr{B}_Y) &= \lim_{n\to\infty} \frac{1}{n} H_\rho \left(\bigvee_{i=0}^{n-1} T^{-i}\xi \times \eta \big| \mathscr{N}_X \times \mathscr{B}_Y \right) \\ &= \lim_{n\to\infty} \frac{1}{n} H_\rho \left(\bigvee_{i=0}^{n-1} T^{-i}\xi \times \{\varnothing, Y\} \big| \mathscr{N}_X \times \mathscr{B}_Y \right) \\ &= \lim_{n\to\infty} \int \frac{1}{n} H_{\mu_y} \left(\bigvee_{i=0}^{n-1} T^{-i}\xi \right) \mathrm{d}\nu(y) \end{aligned}$$

by Lemma 4.6 and equation (4.19). By the dominated convergence theorem, we conclude that

$$h_\rho(T \times I_Y, \xi \times \eta \big| \mathscr{N}_X \times \mathscr{B}_Y) = \int_Y h_{\mu_y}(T, \xi) \, \mathrm{d}\nu(y).$$

Finally, taking sequences $\xi_n \nearrow \mathscr{B}$ and $\eta_n \nearrow \mathscr{B}_Y$, and using Theorem 4.14 and the monotone convergence theorem, we obtain

$$h_\mu(T) = h_\rho(T \times I_Y \,|\, \mathscr{N}_X \times \mathscr{B}_Y) = \int h_{\mu_y}(T)\, \mathrm{d}\nu(y)$$

as claimed. □

Exercises for Section 4.4

Exercise 4.4.1. Prove equation (4.15) in the non-invertible case by establishing the formula $h_\mu(T, \xi) = h_\mu(T, \xi \,|\, \mathscr{E})$ without referring to Section 4.3 (where invertibility was assumed).

Exercise 4.4.2. Prove equation (4.16) for a general convex combination by analyzing the proof of the Abramov–Rokhlin formula (Corollary 4.15) in the case needed, where one of the maps is the identity.

Exercise 4.4.3. Strengthen the variational principle (Theorem 2.22) by showing that if $T : X \to X$ is a continuous map on a compact metric space, then $h_{\mathrm{top}}(T) = \sup_{\mu \in \mathscr{E}^T(X)} h_\mu(T)$.

4.5 An Entropy Calculation

An illuminating example of a compact group automorphism is the map

$$T = T_A : \mathbb{T}^2 \to \mathbb{T}^2$$

defined by

$$T : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} y \\ x + y \end{pmatrix} \pmod 1.$$

This map is associated to the matrix $A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$ in a natural way. Since T is an endomorphism of a compact group, it preserves the Lebesgue measure m on $\mathbb{T}^2$ (see [38, Ex. 2.5]). Alternatively, the invariance of Lebesgue measure follows from the fact that A^{-1} is also an integer matrix and so T_A is invertible and A does not distort area locally (both of these observations follow from the fact that $|\det(A)| = 1$).

In this section we will study (and evaluate) the dynamical entropy of T with respect to the Lebesgue measure, and will discuss what is involved in expressing the entropy of T with respect to other measures.

4.5.1 Entropy of T with respect to the Lebesgue Measure

Theorem 4.24. *The entropy of the automorphism $T = T_A$ of the 2-torus associated to the matrix $A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$ is given by*

$$h_m(T) = \log \rho$$

where $\rho = 1.6\ldots$ is the golden ratio, characterized by $\rho > 1$ and $\rho^2 = \rho + 1$.

Theorem 4.24 is a special case of a general result for automorphisms of the torus, which will be shown in Theorem 3.13 by other methods. The approach taken in this section reveals more of the geometry of the map, and thus tells us more about the entropy of other measures.

We will prove[29] Theorem 4.24 by finding a generator reflecting the geometrical action of T on the torus. This is not the most efficient or general method, but it motivates other ideas presented later. In order to do this, consider first the action of the matrix A on the covering space $\mathbb{R}^2$ of the torus. There are two eigenvectors: $\mathbf{v}^+ = \begin{pmatrix} 1 \\ \rho \end{pmatrix}$, which is dilated by the factor $\rho > 1$, and $\mathbf{v}^- = \begin{pmatrix} 1 \\ -1/\rho \end{pmatrix}$, which is shrunk by the factor $-1/\rho < 0$.

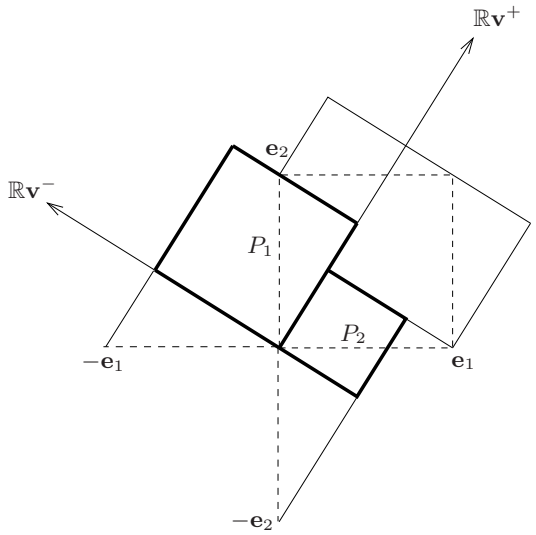

Fig. 4.1. A partition of $\mathbb{T}^2$ adapted to the geometry of the automorphism.

Let $\xi = \{P_1, P_2\}$ denote the partition of $\mathbb{T}^2$ into the two regions shown in Figure 4.1. In Figure 4.1 the square drawn in dashed lines is the unit

square in $\mathbb{R}^2$, which maps under the quotient map $\mathbb{R}^2 \to \mathbb{R}^2/\mathbb{Z}^2$ onto the 2-torus (and the quotient map is injective on the interior of the unit square). The interiors of the bold boxes are the partition elements as labeled, while the thin drawn boxes are integer translates of the two partition elements showing that ξ is genuinely a partition of $\mathbb{T}^2$. Notice that all the sides of these boxes are contained in lines parallel to either $\mathbf{v}^+$, $\mathbf{v}^-$ and going through 0 respectively, $\pm\mathbf{e}_1$ or $\pm\mathbf{e}_2$ (where $\mathbf{e}_1 = (1,0)$ and $\mathbf{e}_2 = (0,1)$). Which element of the partition ξ contains the boundaries of P_1 and P_2 is not specified; since the boundaries are null sets this will not affect the outcome. For now we are only considering the case of the Lebesgue measure m; in Section 4.5.2 other measures and what is needed for this kind of argument will be discussed.

The action of T^{-1} contracts lengths along lines parallel to the expanding eigenvector $\mathbf{v}^+$ for T by a factor of ρ; along lines parallel to the contracting eigenvector $\mathbf{v}^-$, T^{-1} expands by a factor of $-\rho$. Figure 4.2 shows the resulting three rectangles in $\xi \vee T^{-1}\xi$. It is not a general fact that two rectangles in $\mathbb{T}^2$ with parallel sides intersect in a single rectangle, but this happens for all intersections of rectangles in ξ and in $T^{-1}\xi$. Notice that, for example, the rectangle $P_1 \cap T^{-1}P_1$ appears twice on the picture drawn in $\mathbb{R}^2$, but only once in the torus. We suggest that the reader verifies these statements before reading on. For this, note that one can calculate $T^{-1}\xi$ by finding $A^{-1}(\pm\mathbf{e}_i)$ for $i = 1, 2$ and then drawing boxes with sides parallel to $\mathbf{v}^+$ and $\mathbf{v}^-$. We proceed next to show why ξ is such a convenient partition for the map T.

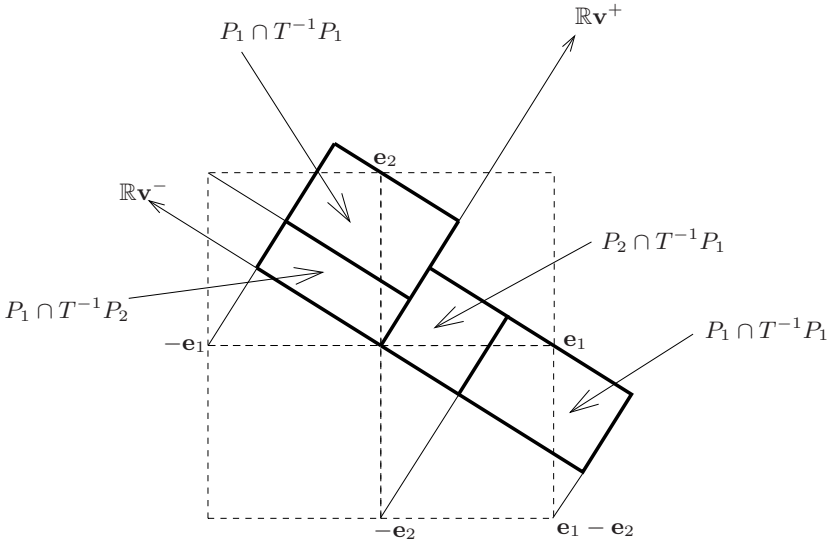

Fig. 4.2. The three rectangles in $\xi \vee T^{-1}\xi$.

Lemma 4.25. *For any $n \geqslant 1$ the elements of the partition*

$$\xi \vee T^{-1}\xi \vee \cdots \vee T^{-n}\xi$$

are rectangles with edges parallel to the eigenvectors. The long side of any such rectangle is parallel to $\mathbf{v}^-$ *with length determined by the element of* ξ *containing it. The short side of any such rectangle is parallel to* $\mathbf{v}^+$ *and has length at most* $2\rho^{-n}$*. Moreover, the atoms for the* σ*-algebra*

$$\mathscr{A} = \xi \vee T^{-1}\xi \vee \cdots = \bigvee_{i=0}^{\infty} T^{-i}\xi \tag{4.20}$$

are line segments parallel to $\mathbf{v}^-$ *as long as the element of* ξ *containing them, and* ξ *is a generator under* T*.*

Proof. We start by proving the first statement by induction. The discussion before the statement of the lemma and Figure 4.2 comprise the case $n = 1$. So assume the statement holds for a given n and consider the partition

$$\eta = T^{-1}\xi \vee \cdots \vee T^{-(n+1))}\xi = T^{-1}\left(\xi \vee \cdots \vee T^{-n}\xi\right).$$

This contains only rectangles with sides parallel to $\mathbf{v}^+$ and $\mathbf{v}^-$ (which will be understood without mention below) which are thinner in the direction of $\mathbf{v}^+$; indeed the maximal thickness has been divided by $\rho > 1$. Along the direction of $\mathbf{v}^-$ they are as long as the element of $T^{-1}\xi$ containing them. Thus

$$\xi \vee \overbrace{T^{-1}\xi \vee \cdots \vee T^{-(n+1)}\xi}^{\eta}$$

contains sets of the form $P \cap Q \subseteq P \cap T^{-1}P'$ for $Q \subseteq T^{-1}P'$, $P, P' \in \xi$, and $Q \in \eta$ (see Figure 4.3).

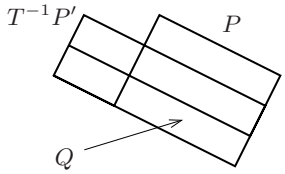

Fig. 4.3. An atom in $\xi \vee T^{-1}\xi \vee \cdots \vee T^{-(n+1)}\xi$.

All of the sets $P, Q, P', T^{-1}P'$ are rectangles, and by assumption Q and $T^{-1}P'$ have the same length in the direction of $\mathbf{v}^-$. Also $P \cap T^{-1}P'$ is again a rectangle whose length along the direction of $\mathbf{v}^-$ is the same as the corresponding length for P (this is the case $n = 1$). Finally, notice that $T^{-1}P'$ is the injective image of a rectangle in $\mathbb{R}^2$. From this we can conclude that

$$P \cap Q = \left(P \cap T^{-1}P'\right) \cap \left(T^{-1}P' \cap Q\right)$$

may be viewed as the image of the intersection of two rectangles in $\mathbb{R}^2$, so $P \cap Q$ is a rectangle. The side of $P \cap Q$ along the direction of $\mathbf{v}^-$ is the intersection of the sides of $P \cap T^{-1}P'$ and $T^{-1}P' \cap Q$, which finishes the induction. We deduce that the σ-algebra $\mathscr{A}$ from equation (4.20) is the σ-algebra comprising measurable sets A with the property that $A \cap P_1$ is a union of lines parallel to $\mathbf{v}^-$ and $A \cap P_2$ is a union of lines parallel to $\mathbf{v}^-$. In other words, $A \in \mathscr{A}$ if $A \cap P_i$ is the direct product of the line segment parallel to $\mathbf{v}^-$ and a Borel subset of the line segment parallel to $\mathbf{v}^+$ in the same way as the rectangle is the direct product of these two line segments. To see this, it is enough to recall that a σ-algebra $\mathscr{C} \subseteq \mathscr{B}$ generated by countably many subintervals $I_n \subseteq I$ of some interval $I \subseteq \mathbb{R}$ coincides with the Borel σ-algebra on I if for every $\varepsilon > 0$ and $x \in I$ there is some I_n containing x with width less than ε. Thus we have shown the statement regarding the atoms of $\mathscr{A}$. Finally, applying T^m to $\mathscr{A}$ shows that the atoms of $T^m\mathscr{A}$ are still parallel to $\mathbf{v}^-$ but have length at most $2\rho^{-m}$. It follows that the atoms for $\bigvee_{i=-\infty}^{\infty} T^{-i}\xi$ are single points, so ξ is a generator. □

By the Kolmogorov–Sinaĭ theorem (Theorem 1.20), Lemma 4.25 reduces the proof of Theorem 4.24 to calculating

$$h_m(T) = h_m(T, \xi) = H_m(T\xi \big| \mathscr{A}).$$

Proof of Theorem 4.24. We wish to compute the conditional information function $I_m(\xi \big| T^{-1}\mathscr{A})$. Using Figure 4.2 and the properties of ξ, we see that

$$I_m(\xi \big| T^{-1}\mathscr{A})(x) = \begin{cases} 0 & \text{for } x \in T^{-1}P_2 \text{ (since then } x \in P_1); \\ \log \rho & \text{for } x \in P_1 \cap T^{-1}P_1; \\ \log \rho^2 & \text{for } x \in T^{-1}P_1 \cap P_2. \end{cases}$$

Here the positive terms are deduced from the fact that the side lengths of the rectangles $T^{-1}P_1$, $P_1 \cap T^{-1}P_1$, and $P_2 \cap T^{-1}P_1$ in the direction of $\mathbf{v}^-$ have the same relative ratios as $1, 1/\rho$, and $1/\rho^2$ respectively. Using the same argument and the fact that T preserves the measure we see that

$$m(P_2) = m(P_2 \cap T^{-1}P_1) = \frac{1}{\rho^2} m(T^{-1}P_1) = \frac{1}{\rho^2} m(P_1),$$

from which, together with the fact that $m(P_1) + m(P_2) = 1$, we can calculate that $m(P_1) = \frac{\rho^2}{1+\rho^2}$. It follows that $m(P_1 \cap T^{-1}P_1) = \frac{\rho}{1+\rho^2}$ and so

$$\begin{aligned} h_m(T, \xi) &= \int_{\mathbb{T}^2} I_m(\xi \big| T^{-1}\mathscr{A})(x) \, \mathrm{d}m(x) \\ &= \frac{1}{\rho} \frac{\rho^2}{1+\rho^2} \cdot \log \rho + \frac{1}{\rho^2} \frac{\rho^2}{1+\rho^2} \cdot 2 \log \rho \\ &= \left(\frac{\rho + 2}{1+\rho^2} \right) \log \rho = \log \rho \text{ since } \rho + 1 = \rho^2. \end{aligned}$$

□

4.5.2 Entropy for Other Invariant Measures

Behind the somewhat enigmatic details of the argument in Section 4.5.1 lies a simple idea reflected in the geometry of the action of T on suitable rectangles: it is contraction (respectively, expansion) along eigenspaces for eigenvalues of absolute value less (resp. greater) than one that contributes to the entropy. However, if we are considering an arbitrary invariant Borel measure μ, naturally the properties of the measure also play a role in computing $h_\mu(T)$. In this[30] section, we will explain why $h_\mu(T)$ depends mainly on the properties of the conditional measures $\mu_x^{\mathscr{A}}$ as x varies in $\mathbb{T}^2$. Here $\mathscr{A}$ is again defined as in Lemma 4.25 using the partition ξ. Assuming that μ gives zero mass to the origin, the boundaries of the elements of ξ are null sets and nothing changes in Lemma 4.25. To see this, recall that the boundaries of these rectangles are made of pieces of $\mathbb{R}\mathbf{v}^+$ and $\mathbb{R}\mathbf{v}^-$ through integer points. In $\mathbb{T}^2$ this means that these points either approach the origin in their backward orbit or in their forward orbit. In either case we can apply Poincaré recurrence (see [38, Th. 2.11]) to see that the boundary is a μ-null set: If K is the complement of a neighborhood of 0, points in $K \cap \mathbb{R}\mathbf{v}^-$ are non-returning to K after finitely many steps. Define

$$U^- = \mathbb{R}\mathbf{v}^-,$$

the *stable subgroup* for T and

$$U^+ = \mathbb{R}\mathbf{v}^+,$$

the *unstable subgroup* of T. For $x \in \mathbb{R}^2$ the *stable manifold* through x is the coset $x + U^-$ and the *unstable manifold* is $x + U^+$. Finally, for $\delta > 0$ we let

$$B_\delta^{U^-}(x) = x + \big(U^- \cap B_\delta(0)\big)$$

denote the δ-neighborhood of x inside the stable manifold. The δ-neighborhood of x inside the unstable manifold, $B_\delta^{U^+}(x)$, is defined similarly. It is important to note that we consider the intersections like $U^- \cap B_\delta(0)$ in the covering space $\mathbb{R}^2$, while the translation by the point x is made in $\mathbb{T}^2$ to define finally a subset of $\mathbb{T}^2$.

Theorem 4.26. *Let $T = T_A$ be the automorphism of the 2-torus associated to the matrix $A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$ and let μ be a T-invariant non-atomic probability measure on $\mathbb{T}^2$. Then for μ-almost every x we have*

$$h_{\mu_x^{\mathscr{E}}}(T) = \lim_{n\to\infty} \frac{-\log \mu_x^{\mathscr{A}}[x]_{T^n\mathscr{A}}}{n}. \tag{4.21}$$

That is, the limit on the right-hand side of equation (4.21) exists and equals the entropy of T with respect to the ergodic component $\mu_x^{\mathscr{E}}$ of μ at x. In particular, by Theorem 4.23,

$$h_\mu(T) = \int h_{\mu_x^{\mathscr{E}}}(T)\,\mathrm{d}\mu(x) = \int \lim_{n\to\infty} \frac{-\log \mu_x^{\mathscr{A}}[x]_{T^n\mathscr{A}}}{n}\,\mathrm{d}\mu(x).$$

Proof of Theorem 4.26. Define $f(x) = I_\mu(\xi\big|T^{-1}\mathscr{A})(x)$ for the generating partition ξ as in Lemma 4.25, and let $\mathscr{A}$ be as in Lemma 4.25. Then

$$\begin{aligned} f(T^{-1}x) &= I_\mu(\xi\big|T^{-1}\mathscr{A})(T^{-1}x) \\ &= I_\mu(T\xi\big|\mathscr{A})(x) \end{aligned}$$

and so on:

$$f(T^{-k}x) = I_\mu(T^k\xi\big|T^{k-1}\mathscr{A})(x).$$

Thus

$$\begin{aligned} \frac{1}{n}\sum_{k=0}^{n-1} f(T^{-k}x) &= \frac{1}{n}\sum_{k=0}^{n-1} I_\mu(T^k\xi\big|T^{k-1}\mathscr{A})(x) \\ &= \frac{1}{n} I_\mu(\xi\vee T\xi\vee\cdots\vee T^{n-1}\xi\big|\mathscr{A})(x) \\ &= -\frac{n-1}{n}\cdot\frac{1}{n-1}\log\mu_x^{\mathscr{A}}\left([x]_{T^{n-1}\mathscr{A}}\right), \end{aligned}$$

showing the limit of the expression on the right of equation (4.21) exists and is equal to $E\left(I_\mu(\xi\big|T^{-1}\mathscr{A})\big|\mathscr{E}\right)(x)$ (notice that $\mathscr{E}\subseteq\mathscr{A}$ by Theorem 4.19), which in turn may be seen to be $h_{\mu_x^{\mathscr{E}}}(T,\xi) = h_{\mu_x^{\mathscr{E}}}(T)$ by integrating the equation

$$\begin{aligned} I_\mu\left(\xi\big|T^{-1}\mathscr{A}\right)(y) &= -\log\mu_y^{T^{-1}\mathscr{A}}([y]_\xi) \\ &= -\log\left(\mu_x^{\mathscr{E}}\right)_y^{T^{-1}\mathscr{A}}([y]_\xi) \\ &= I_{\mu_x^{\mathscr{E}}}\left(\xi\big|T^{-1}\mathscr{A}\right)(y) \end{aligned}$$

over y. □

The dependence in Theorem 4.26 on the geometry of $[x]_{T^n\mathscr{A}}$ is slightly unsatisfactory. We know that $[x]_{T^n\mathscr{A}}$ is an interval in the set $x+U^-$ of size comparable to ρ^{-n}, but what we do not know is the position of x within that interval. Assuming for the moment that this does not have any influence on the entropy, we expect that the limit $h_x = s_x\log\rho$, where

$$s_x = \lim_{\delta\to 0}\frac{\log\mu_x^{\mathscr{A}}(B_\delta^{U^-}(x))}{\log\delta},$$

exists, and we may call s_x the *local dimension* of μ along the stable manifold of T. Moreover, $h_\mu(T)$ is then given by the logarithm of the contraction factor times the average dimension of μ along the stable manifold*. We will prove

* Because of the symmetry $h_\mu(T) = h_\mu(T^{-1})$ the same argument gives a similar result phrased in terms of the unstable manifolds.

this extension in greater generality later, but only for a specific collection of algebraic dynamical systems.

To justify the terminology "dimension" for s_x we show next that in our setting of one-dimensional stable manifolds it automatically lies between 0 and 1. This quantity also behaves like a dimension in that it gives the power of the decay rate of the measure of balls.

Lemma 4.27. *For any finite measure ν on the Borel sets in $[0,1]$ with the property that*

$$g(t) = \lim_{\delta \to 0} \frac{\log \nu(B_\delta(t))}{\log \delta}$$

exists for almost every $t \in [0,1]$, $g(t) \leqslant 1$ almost everywhere.

Proof. Fix $\varepsilon > 0$, and let $A = \{t \in [0,1] \mid g(t) > 1 + 2\varepsilon\}$. It is enough to show that $\nu(A) \to 0$ as $\varepsilon \to 0$. Write

$$A_k = \{t \mid \nu\left(B_{2^{-k}}(t)\right) > 2^{-k(1+\varepsilon)}\}$$

so that

$$A \subseteq \bigcap_{\ell \geqslant 1} \bigcup_{k \geqslant \ell} A_k.$$

If $I \subseteq [0,1]$ is an interval of length 2^{-k-1} with $A_k \cap I \neq \varnothing$ then

$$\nu(I) < 2^{-k(1+\varepsilon)},$$

so

$$\nu(A_k) \leqslant 2^{k+1} 2^{-k(1+\varepsilon)} = 2\left(\frac{2}{2^{1+\varepsilon}}\right)^k$$

and therefore

$$\nu\left(\bigcup_{k \geqslant \ell} A_k\right) \leqslant 2 \sum_{k \geqslant \ell} \left(\frac{2}{2^{1+\varepsilon}}\right)^k = \left(\frac{2}{2^{1+\varepsilon}}\right)^\ell \cdot \frac{2}{1 - 2^{-\varepsilon}} \to 0$$

as $\ell \to \infty$. It follows that $\nu(A) = 0$. □

4.5.3 Maximality Property of Lebesgue Measure

In this section we will use our detailed knowledge of the geometry of the action of T_A on $\mathbb{T}^2$ to prove the following theorem, which gives weight to our earlier claim on p. 8 that certain natural invariant measures can be characterized as being those with maximal entropy[(31)].

Theorem 4.28. *Let $T = T_A$ be the automorphism of the 2-torus associated to the matrix $A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$, and let μ be a T-invariant probability measure on $\mathbb{T}^2$. Then $h_\mu(T) \leqslant \log \rho = h_m(T)$, and equality holds if and only if $\mu = m$.*

Thus no T-invariant Borel measure can give more entropy than Lebesgue measure. In Section 2.3 this maximality property will be related to properties of the map T viewed as a continuous map. Before proving the theorem, we show how a pair of mutually singular measures are widely separated by a sequence of generating partitions.

Lemma 4.29. *Let $(X, \mathscr{B})$ be a Borel space and let μ_1, μ_2 be two Borel probability measures that are singular with respect to each other. If (ξ_n) is an increasing sequence of $\mathscr{B}$-measurable partitions which generate $\mathscr{B}$ modulo the measure $\lambda = \mu_1 + \mu_2$, then there exists a sequence of sets $A_n \in \sigma(\xi_n)$ with the property that $\mu_1(A_n) \to 0$ and $\mu_2(A_n) \to 1$ as $n \to \infty$.*

Proof. Since $\lambda = \mu_1 + \mu_2$ and $\mu_1 \perp \mu_2$ there is a set $B \in \mathscr{B}$ with

$$\frac{\mathrm{d}\mu_1}{\mathrm{d}\lambda} = \chi_B$$

and

$$\frac{\mathrm{d}\mu_2}{\mathrm{d}\lambda} = \chi_{B^c}$$

(see [38, Sect. C.4]). We will apply results from [38, Chap. 5] to λ (λ is not a probability measure, but the results needed apply to finite measures simply by normalizing the measure to be a probability measure). By assumption,

$$\mathscr{A}_n = \sigma(\xi_n) \nearrow \mathscr{B} \pmod{\lambda},$$

so by the increasing martingale theorem (see [38, Th. 5.5]),

$$f_n = E_\lambda\left(\chi_B \middle| \mathscr{A}_n\right) \longrightarrow \chi_B$$

in L^1_λ. We define

$$A_n = \{x \mid f_n(x) > \tfrac{1}{2}\} \in \mathscr{A}_n.$$

Then

$$\begin{aligned}\|\chi_B - \chi_{A_n}\|_{L^1_\lambda} &= \int_B (1 - \chi_{A_n})\,\mathrm{d}\lambda + \int_{B^c} \chi_{A_n}\,\mathrm{d}\lambda \\ &= \mu_1(A_n^c) + \mu_2(A_n)\end{aligned}$$

and

$$\|\chi_B - \chi_{A_n}\|_{L^1_\lambda} = \lambda\left(\{x \in B \mid f_n(x) \leqslant \tfrac{1}{2}\} \cup \{x \notin B \mid f_n(x) > \tfrac{1}{2}\}\right) \longrightarrow 0$$

as $n \to \infty$, which proves the lemma. □

We first prove Theorem 4.28 using an approach due to Adler and Weiss [4].

First proof of Theorem 4.28. By Theorem 4.23 we may assume without loss of generality that μ is ergodic: if all ergodic components have entropy no more than $\log \rho$, then the same holds for μ, with equality if and only if there

is equality for almost every ergodic component. Since $\{0\}$ is a T-invariant set, we have $\mu(\{0\}) \in \{0,1\}$, and the point measure δ_0 supported on $\{0\}$ clearly has zero entropy. Thus we may assume that $\mu(\{0\}) = 0$ which, as discussed in Section 4.5.2, allows us to use the generator ξ described in Section 4.5.1*.

We are assuming that $h_\mu(T) \geqslant \log\rho$; if $\mu \neq m$ then, by [38, Lem. 4.6], μ and m are mutually singular. Hence, by Lemma 4.29, there is a sequence of sets $A_n \in \sigma\left(\bigvee_{i=0}^{n-1} T^{-n}\xi\right)$ with $m(A_n) \to 0$ and $\mu(A_n) \to 1$ as $n \to \infty$. Write $N_n(B)$ for the number of elements of the partition $\bigvee_{i=0}^{n-1} T^{-n}\xi$ required to cover $B \in \sigma\left(\bigvee_{i=0}^{n-1} T^{-n}\xi\right)$.

From the geometry of the partition $\xi_n = \bigvee_{i=0}^{n-1} T^{-n}\xi$ (as described in Section 4.5.1) we can see that all elements of the partition ξ_n have Lebesgue measure at least $\frac{C}{\rho^n}$ for some constant $C < 1$. It follows that

$$m(B) \geqslant C\frac{N_n(B)}{\rho^n} \tag{4.22}$$

for any $B \in \sigma(\xi_n)$. Since $h_\mu(T) = \inf_{n\in\mathbb{N}}\{\frac{1}{n}H_\mu(\xi_n)\} \geqslant \log\rho$ by assumption, the additivity of the entropy function gives

$$\log\rho^n \leqslant H_\mu(\xi_n) = H_\mu\left(\xi_n \middle| \{A_n, A_n^c\}\right) + H_\mu\left(\{A_n, A_n^c\}\right).$$

Using Lemma 4.6 and the fact that entropy is maximized if all partition elements have the same measure (Proposition 1.5), we get

$$\begin{aligned}\log\rho^n &\leqslant \mu(A_n)\log N_n(A_n) + \mu(A_n^c)\log N_n(A_n^c) + \log 2\\ &\leqslant \mu(A_n)\log\left(m(A_n)\rho^n\right) + \mu(A_n^c)\log\left(m(A_n^c)\rho^n\right) + \log 2 - 2\log C\end{aligned}$$

by using the inequality (4.22). Subtracting

$$\log\rho^n = \mu(A_n)\log\rho^n + \mu(A_n^c)\log\rho^n$$

from both sides gives

$$0 \leqslant \mu(A_n)\log\left(m(A_n)\right) + \mu(A_n^c)\log\left(m(A_n^c)\right) + \log 2 - 2\log C. \tag{4.23}$$

However, by Lemma 4.29 we have $m(A_n) \to 0$ and $\mu(A_n) \to 1$ as $n \to \infty$, so the right-hand side of equation (4.23) approaches $-\infty$ as $n \to \infty$. This gives a contradiction to the assumptions $\mu \neq m$ and $h_\mu(T) \geqslant \log\rho$ and concludes the proof. □

A different proof of Theorem 4.28 may be given using a method of Margulis and Tomanov [93, Sec. 9], which generalizes to settings beyond group automorphisms.

* To be precise, we have not really defined a partition until we specify which atoms contain the various edges. Assigning edges to atoms in any reasonable fashion makes ξ a generator. If $\mu(\{0\}) = 0$, then it is safe to ignore the edges.

Second proof of Theorem 4.28. As discussed above, we may assume that $\mu(\{0\}) = 0$. We will demonstrate the unique maximality using strict convexity of $x \mapsto -\log x$, and by using the functions $x \mapsto I_\mu\left(\xi \middle| T^{-1}\mathscr{A}\right)(x)$ and

$$f(x) = \begin{cases} 0 & \text{for } x \in T^{-1}P_2; \\ \log \rho & \text{for } x \in P_1 \cap T^{-1}P_1; \\ \log \rho^2 & \text{for } x \in P_2 \cap T^{-1}P_1 \end{cases}$$

to compare μ to m.

Recall from the start of the proof of Theorem 4.24 on p. 122 that $f(x) = I_m(\xi|T^{-1}\mathscr{A})(x)$, but notice that the definition of f here does not rely on m, and defines a measurable function on all of $\mathbb{T}^2$.

We first calculate $\int f \, \mathrm{d}\mu$ using the function

$$h(x) = \begin{cases} \log \rho & \text{for } x \in P_1; \\ 0 & \text{for } x \in P_2, \end{cases}$$

which is useful because

$$f(x) = \log \rho + h(Tx) - h(x),$$

so that $\int f \, \mathrm{d}\mu = \log \rho$. As a partial explanation for the choice of the function h, we note that $e^{h(x)}$ coincides (up to a constant) with the length of the rectangle $P \in \xi$ that contains x, in the direction of $\mathbf{v}^-$.

Now we compute

$$\int \left(f - I_\mu(\xi|T^{-1}\mathscr{A})\right)(y) \, \mathrm{d}\mu_x^{T^{-1}\mathscr{A}}(y) = \int \left(-\log\left(\frac{\mathrm{e}^{-f(y)}}{\mu_y^{T^{-1}\mathscr{A}}([y]_\xi)}\right)\right) \mathrm{d}\mu_x^{T^{-1}\mathscr{A}} \tag{4.24}$$

as follows. For $x \in T^{-1}P_2$, this is 0. For $x \in T^{-1}P_1$ we find that equation (4.24) is equal to

$$-\log\left(\frac{1/\rho}{\mu_x^{T^{-1}\mathscr{A}}(P_1)}\right)\mu_x^{T^{-1}\mathscr{A}}(P_1) - \log\left(\frac{1/\rho^2}{\mu_x^{T^{-1}\mathscr{A}}(P_2)}\right)\mu_x^{T^{-1}\mathscr{A}}(P_2).$$

Assuming that $x \in T^{-1}P_1$, the convexity of $x \mapsto -\log x$ (Lemma 1.4) therefore shows that

$$\int \left(f - I_\mu(\xi|T^{-1}\mathscr{A})\right) \mathrm{d}\mu_x^{T^{-1}\mathscr{A}} \geqslant -\log\left(\frac{1/\rho}{\mu_x^{T^{-1}\mathscr{A}}(P_1)}\mu_x^{T^{-1}\mathscr{A}}(P_1) + \frac{1/\rho^2}{\mu_x^{T^{-1}\mathscr{A}}(P_2)}\mu_x^{T^{-1}\mathscr{A}}(P_2)\right) = 0.$$

The same inequality holds trivially for $x \in T^{-1}P_2$. Integrating with respect to μ over $x \in \mathbb{T}^2$ gives

$$\log \rho - h_\mu(T) \geqslant 0.$$

Assume now that $h_\mu(T) = \log\rho$. Then by strict convexity of $x \mapsto -\log x$ (Lemma 1.4), we see that

$$\frac{1/\rho}{\mu_x^{T^{-1}\mathscr{A}}(P_1)} = \frac{1/\rho^2}{\mu_x^{T^{-1}\mathscr{A}}(P_2)} = 1$$

for μ-almost everywhere $x \in T^{-1}(P_1)$.

Combined these show that

$$I_\mu\left(\xi\middle|T^{-1}\mathscr{A}\right) = f(x) \tag{4.25}$$

μ-almost everywhere. Hence

$$\begin{aligned} I_\mu\left(T^n\xi \vee T^{n-1}\xi \vee \cdots \vee \xi\middle|T^{-1}\xi\right)(x) = {} & I_\mu\left(T^n\xi\middle|T^n\xi \vee \cdots \vee \xi \vee T^{-1}\mathscr{A}\right)(x) \\ & + I_\mu\left(T^n\xi\middle|T^n\xi \vee \cdots \vee \xi \vee T^{-1}\mathscr{A}\right)(x) \\ & + \cdots + I_\mu\left(\xi\middle|T^{-1}\mathscr{A}\right)(x) \end{aligned}$$

by Proposition 4.9(1). Thus

$$I_\mu\left(T^n\xi \vee \cdots \vee\middle|T^{-1}\xi\right)(x) = \sum_{k=0}^{n} f(T^{-k}x). \tag{4.26}$$

As $f(x)$ is an everywhere-defined choice of $I_m\left(\xi\middle|T^{-1}\mathscr{A}\right)(x)$ for μ-almost every x, equation (4.26) just means that $\mu_x^{T^{-1}\mathscr{A}}$ gives the atoms of

$$T^n\mathscr{A} = T^n\xi \vee \cdots \vee \xi \vee T^{-1}\mathscr{A}$$

the same weight as the normalized Lebesgue measure on the line segment

$$[x]_{T^{-1}\mathscr{A}} \subseteq x + \mathbb{R}\mathbf{v}^{-1}$$

would.

However, as n increases the atoms of $T^n\mathscr{A}$ become arbitrarily small subintervals of the line segment $[x]_{T^{-1}\mathscr{A}}$, so that the above argument shows that $\mu_x^{T^{-1}\mathscr{A}}$ coincides with the normalized Lebesgue measure on the line segment $[x]_{T^{-1}\mathscr{A}}$ for $x \in Y$; equivalently the same property holds for $\mathscr{A} = \bigvee_{n=0}^{\infty} T^{-n}\xi$.

By symmetry of entropy (Proposition 1.17) we have $h_\mu(T) = h_\mu(T^{-1})$, and a similar argument applies to the atoms of the σ-algebra $\bigvee_{n=0}^{\infty} T^n\xi$, which are line segments in the direction of $\mathbb{R}\mathbf{v}^+$ which are as long as the element $P \in \xi$ which contains x, and this holds for x in a subset $Z \subseteq Y$ of full measure as illustrated in Figure 4.4.

Fix some $P \in \xi$, and recall that P is a rectangle so that we may think of P as a product $I^+ \times I^-$ of an interval $I^+ \subseteq \mathbb{R}\mathbf{v}^+$ and an interval $I^- \subseteq \mathbb{R}\mathbf{v}^-$. In this coordinate system $\mathscr{A}|_P$ corresponds to the product $\mathscr{B}_{I^+} \times \mathscr{N}_{I^-}$ of the Borel σ-algebra $\mathscr{B}_{I^+}$ on I^+ and the trivial σ-algebra $\mathscr{N}_{I^-} = \{\varnothing, I^-\}$

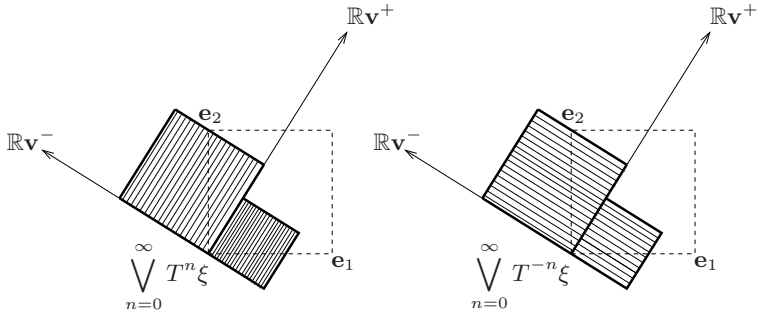

Fig. 4.4. Identifying the conditional measures on the two one-sided orbits of ξ to be one-dimensional Lebesgue measures.

on I^-, and we have shown that the conditional measures for $\mathscr{B}_{I^+} \times \mathscr{N}_{I^-}$ are of the form $\delta_s \times m_{I^-}$, where m_{I^-} denotes the normalized Lebesgue measure on I^-. If λ_+ denotes the projection of the measure on $I^+ \times I^-$ to the I^+-coordinate (so that $\lambda_+(B)$ is the measure of $B \times I^-$) then the argument above implies that the measure $\mu|_P$ has the form $\lambda_+ \times m_{I^-}$ in the coordinate system $I^+ \times I^-$ of P. However, by the same argument (using the information obtained from $\bigvee_{n=0}^{\infty} T^n\xi$), we know that μ_P also has the form $m_{I^+} \times \lambda_-$ for some measure λ_- on I^-. Together this implies that indeed $\mu|_P$ is equal to $\mu(P)$ times the normalized two-dimensional Lebesgue measure on the rectangle P. Therefore

$$\mu = \frac{\mu(P_1)}{m(P_1)} m|_{P_1} + \frac{\mu(P_2)}{m(P_2)} m|_{P_2}.$$

Now $T^{-1}(P_2) \subseteq P_1$, so that we have

$$\mu(P_2) = \frac{\mu(P_2)}{m(P_2)} m(P_2) = \mu(T^{-1}(P_2)) = \frac{\mu(P_1)}{m(P_1)} m(T^{-1}(P_2)) = \frac{\mu(P_1)}{m(P_1)} m(P_2),$$

and hence $\frac{\mu(P_1)}{m(P_1)} = \frac{\mu(P_2)}{m(P_2)}$. It follows that $\mu = m$. □

Exercises for Section 4.5

Exercise 4.5.1. Generalize the entropy calculation from Section 4.5 to the automorphism associated to any matrix $A \in \mathrm{GL}_2(\mathbb{Z})$ with an eigenvalue of absolute value exceeding one, and generalize the proof of Theorem 4.28 to show that Lebesgue measure is the only measure with this entropy.

Notes to Chapter 4

[25] (Page 100) Notice that this equivalent formulation only applies to countably-generated σ-algebras; the more sophisticated definition given on [38, p. 130] applies in general.

[26] (Page 108) The main result concerning the existence of generators is due to Krieger [65]: if $(X, \mathscr{B}, \mu, T)$ has finite entropy, then a generator exists with d atoms, where $\mathrm{e}^{h(T)} \leqslant d \leqslant \mathrm{e}^{h(T)} + 1$. Notice that by Proposition 1.5 and Proposition 1.16(1) it is not possible for there to be a generator with fewer atoms, so this result is optimal. The Krieger generator theorem is shown in the books of Rudolph [125] and Parry [112].

[27] (Page 109) The map T may be thought of as a skew-product construction, and the entropy formula is proved by Abramov and Rokhlin [2]. It is generalized to actions of countable amenable groups by Ward and Zhang [143].

[28] (Page 113) The converse is not true: there are measure-preserving systems with trivial Pinkser algebra that are not isomorphic to Bernoulli shifts. The distinction is a subtle one, and erroneous arguments that the two properties are the same were put forward by Wiener [147] among others. An uncountable family of non-isomorphic measure-preserving transformations with trivial Pinsker algebra, none of which is isomorphic to a Bernoulli shift, is constructed by Ornstein and Shields [106]. Smooth examples of this sort were constructed by Katok [69].

[29] (Page 119) These geometrically natural generators were introduced in work of Adler and Weiss [4], [5].

[30] (Page 123) The material in this section is a very special case of the beginning of a profound theory of the entropy of diffeomorphisms developed by many researchers including Pesin [114], Ledrappier and Young [80], [81] and Mañé [92].

[31] (Page 125) This is a special case of a general result due to Berg [9]: Haar measure is the unique maximal measure for an ergodic automorphism of a compact group with finite entropy (finite entropy is clearly required: for example, the full shift $\prod_{-\infty}^{\infty} \mathbb{T}$ with alphabet the circle has infinite topological entropy, and any measure μ on $\mathbb{T}$ with the property that for any K there is a partition ξ of $\mathbb{T}$ with $H_\mu(\xi) \geqslant K$ defines a measure $\mu^\infty = \prod_{-\infty}^{\infty} \mu$ that is shift-invariant and has maximal entropy). Lemma 2.30 is also a special case of Berg's theorem, since the full shift on s symbols may be viewed as an automorphism of the compact group $\prod_{-\infty}^{\infty} \mathbb{Z}/s\mathbb{Z}$. Berg's theorem extends to actions of $\mathbb{Z}^d$ by automorphisms of a compact group under an additional hypothesis: Lind, Schmidt and Ward [87] show that for such an action with finite entropy, Haar measure is the unique measure of maximal entropy if and only if the system has completely positive entropy (which is equivalent to ergodicity for the case $d = 1$ of a single automorphism).

5

Commuting Automorphisms

Automorphisms or endomorphisms of (infinite) compact groups are soft in the following sense: there are many invariant probability measures, and many closed invariant subsets. We have already seen the symbolic coding of the map $x \mapsto 2x \pmod 1$ and of the toral automorphism corresponding to the matrix $\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$; in each case the symbolic description allows many invariant measures and closed invariant sets to be found. Furstenberg [46] noted that the situation is very different for measures or closed invariant sets invariant under two genuinely distinct endomorphisms. We being this chapter with his original topological result for closed subsets of the circle invariant under two endomorphisms, and go on to describe two related measurable results. One of these constrains the possible joinings between two different algebraic $\mathbb{Z}^2$-actions, and one constrains probability measures invariant under $x \mapsto 2x \pmod 1$ and $x \mapsto 3x \pmod 1$.

5.1 Closed Invariant Sets: Furstenberg's Theorem

Before addressing the measurable questions, we describe[(32)] a simple topological analog. The properties being dealt with in this chapter concern the semigroup $\{2, 3, 4, 6, 9, 12, \dots\} \subseteq \mathbb{N}$ generated by 2 and 3, and we begin with some general observations about such semigroups.

Definition 5.1. *A multiplicative semigroup $S \subseteq \mathbb{N}$ is* lacunary *if there is some $a \in \mathbb{N}$ with the property that any $s \in S$ is a power of a.*

Clearly the semigroup generated by a single element is lacunary; for example $\{1, 2, 4, 8, \dots\}$ is lacunary. There are many others however: the semigroup $\{4, 8, 16, 32, \dots\}$ is lacunary but not generated by any element.

Each $k \in \mathbb{N}$ defines an endomorphism $S_k : \mathbb{T} \to \mathbb{T}$ defined by $S_k(t) = kt \pmod 1$. A set $A \subseteq \mathbb{T}$ is called S_k-invariant if $S_k x \in A$ whenever $x \in A$, and

is called S-invariant for a subset $S \subseteq \mathbb{N}$ if $S_k x \in A$ whenever $x \in A$ and $k \in S$. Lacunary semigroups have many non-trivial closed invariant sets, as shown in the next example.

Example 5.2. The middle-third Cantor set

$$\left\{ x \in \mathbb{T} \mid x = \sum_{n=1}^{\infty} e_n 3^{-n} \text{ has } e_n \in \{0,2\} \text{ for all } n \geqslant 1 \right\}$$

is invariant under S_3 (and hence under any semigroup in the lacunary semigroup $\{3, 9, 27, \dots\}$).

The next result, due to Furstenberg [46], shows that there are no non-trivial closed invariant sets under the action of a non-lacunary semigroup. This short proof is taken from a paper of Boshernitzan [12].

Theorem 5.3 (Furstenberg). *Let S be a non-lacunary semigroup in $\mathbb{N}$ and let A be a closed subset of $\mathbb{T}$ invariant under S. Then either A is finite or $A = \mathbb{T}$.*

Two elements s_1, s_2 of a semigroup are said to be *multiplicatively independent* if $s_1^m = s_2^n$ for $m, n \in \mathbb{N}_0$ implies that $m = n = 0$.

Lemma 5.4. *The following properties of a semigroup $S \subseteq \mathbb{N}$ are equivalent.*

(1) *S is non-lacunary;*
(2) *S contains two multiplicatively independent elements;*
(3) *if $S = \{s_1, s_2, \dots\}$ with $s_1 < s_2 < s_3 < \cdots$ then $\frac{s_{n+1}}{s_n} \to 1$ as $n \to \infty$.*

Proof. The equivalence of (1) and (2) is clear. Now $\log S$ is an additive semigroup in $\mathbb{R}$, so $L = \log S - \log S$ is an additive subgroup of $\mathbb{R}$, and is therefore either dense or discrete.

If L is discrete, then $L \subseteq \mathbb{Z}\ell$ for some $\ell > 0$, so $\log S \subseteq \mathbb{Z}\ell + \log s$ for all $s \in S$; in particular $2 \log s = n\ell + \log s$ so $\log s \in \mathbb{Z}\ell$, which shows that $\log S$ is discrete and $S \subseteq (\exp(\ell))^{\mathbb{N}}$ is lacunary.

If L is dense, we may assume that $\log S$ is generated by $\{\ell_1, \ell_2, \dots\}$, so

$$L = \bigcup_{n=1}^{\infty} \left(\log S - n(\ell_1 + \cdots + \ell_n)\right),$$

with the sets on the right-hand side increasing. Fix $\varepsilon > 0$, and suppose that the sets $\log S - n(\ell_1 + \cdots + \ell_n)$ omitted an interval of length ε in $(0, \infty)$. Then, for a suitable choice of k_n, the set

$$\log S - n(\ell_1 + \cdots + \ell_n) - k_n \ell_1$$

must omit an interval of length ε somewhere in $(-\ell_1, 0)$. Now

$$\log S - n(\ell_1 + \cdots + \ell_n) - k_n\ell_1 \supseteq \log S - n(\ell_1 + \cdots + \ell_n),$$

which contradicts the fact that the sets $\log S - n(\ell_1 + \cdots + \ell_n)$ become dense. It follows that there is some N for which $n > N$ implies that

$$\log S - n(\ell_1 + \cdots + \ell_n)$$

is ε-dense in $(0, \infty)$, so $\log S$ is ε-dense in $(n(\ell_1 + \cdots + \ell_n), \infty)$.

Thus there is the following dichotomy. If L is discrete then $\log S$ is contained in a discrete subgroup, so any two elements of S must be multiplicatively dependent and S is lacunary. If L is dense then $\log S$ becomes more and more dense towards infinity, showing property (3). □

Proof of Theorem 5.3. Write A' for the set of limit points of A and assume that A is infinite, so A' is a non-empty closed invariant set. We claim that it must contain a rational point. Assume for the purposes of a contradiction that A' does not contain any rational, and fix $\varepsilon > 0$. Since S is non-lacunary, we may choose multiplicatively independent numbers $p, q \in S$. Find $t \geqslant 3$ with the properties that $t\varepsilon > 1$ and $\gcd(p, t) = \gcd(q, t) = 1$. It follows that

$$p^u \equiv q^u \equiv 1 \pmod{t} \tag{5.1}$$

where $u = \phi(t)$ (ϕ is the Euler function). Define a sequence of sets

$$B_{t-1} \subseteq B_{t-2} \subseteq \cdots \subseteq B_0 = A'$$

by

$$B_{j+1} = \{x \in B_j \mid x + \tfrac{1}{t} \in B_j \pmod 1\} \tag{5.2}$$

for each j, $0 \leqslant j \leqslant t-2$. We prove the following statements by induction:

- B_j is invariant under S_{p^u} and S_{q^u}.
- B_j is a closed infinite set of irrational numbers.

For $j = 0$ both properties hold by assumption; assume they hold for some j, $0 \leqslant j \leqslant t-2$. Define a set $D_j = B_j - B_j$. Since B_j is compact by assumption, D_j is closed; since B_j is invariant under both S_{p^u} and S_{q^u}, so is D_j; finally 0 must be a limit point of D_j since B_j is infinite. By assumption, the semigroup S' generated by p^u and q^u is non-lacunary, so by Lemma 5.4(3) the elements $s_1 < s_2 < \cdots$ of that semigroup have $\frac{s_{n+1}}{s_n} \to 1$ as $n \to \infty$. We claim that $D_j = \mathbb{T}$. Fix $\delta > 0$ and choose N so that $\frac{s_{n+1}}{s_n} < 1 + \delta$ for $n > N$. Since 0 is a limit point of D_j, we may find $x_n \in D_j$ with $0 \neq |x_n| < \delta/s_n$, then the finite set $\{sx_n \mid s \in S', s_n \leqslant s \leqslant 1/|x_n|\}$ is δ-dense and lies in D_j. It follows that $D_j = \mathbb{T}$.

We deduce that B_{j+1} is non-empty. By the choice of u and equation (5.1), the set B_{j+1} is invariant under S_{p^u} and S_{q^u} and is therefore infinite. Finally, B_{j+1} is a closed set because B_j is a closed set and the condition in equation (5.2) is closed.

We deduce by induction that each of the sets B_j is non-empty, and in particular B_{t-1} is non-empty. Pick any point $x_0 \in B_{t-1}$ and write $x_i = x_0 + \frac{i}{t}$ for $0 \leqslant i \leqslant t-1$. By choice of t the set $C = \{x_i \mid 0 \leqslant i \leqslant t-1\}$ is ε-dense in $\mathbb{T}$ and $C \subseteq B_0 = A'$. Since ε was arbitrary, it follows that A' is dense in $\mathbb{T}$, contradicting the assumption that A' does not contain any rationals.

Thus A' contains some rational $r = n/t$ say. Recall that p and q are multiplicatively independent elements of S. We may assume (replacing r by $p^a q^b r$ for suitable a, b if need be) that

$$\gcd(n,t) = \gcd(p,t) = \gcd(q,t) = 1.$$

As before, choose $u = \phi(t)$, so that $p^u \equiv q^u \equiv 1 \pmod{t}$. The sets A and A' are both invariant under S_{p^u} and S_{q^u}, and by choice of u so are their translates $A' - r$ and $A - r$.

Now 0 lies in $A' - r$; in the argument above concerning D_j we showed that this is enough to show that $A - r = \mathbb{T}$, and therefore $A = \mathbb{T}$. □

5.2 Joinings

Recall that a joining between two measure-preserving systems $(X, \mathscr{B}_X, \mu, T)$ and $(Y, \mathscr{B}_Y, \nu, S)$ is a probability measure ρ on the product space

$$(X \times Y, \mathscr{B}_x \otimes \mathscr{B}_Y)$$

invariant under the product map $T \times S$ and with the property that

$$\rho(A \times Y) = \mu(A), \rho(X \times B) = \nu(B)$$

for any $A \in \mathscr{B}_X, B \in \mathscr{B}_Y$. The space of joinings between two ergodic group automorphisms is a vast and unmanageable collection in general, whereas the space of joinings between two ergodic circle rotations is much smaller. One of the manifestations of rigidity for commuting automorphisms is that there are very few joinings[(33)] of mixing algebraic $\mathbb{Z}^d$-actions with the property that individual elements act with finite entropy for $d \geqslant 2$. In this section we record a particularly simple instance of this phenomena on disconnected groups [37], which also helps to motivate some arguments that will appear in the proof of Rudolph's theorem (Theorem 5.8).

Notice that the relatively independent joining (see [38, Def. 6.15]) means that disjointness between two systems implies that they have no non-trivial common factors.

Recall Ledrappier's example, which is the $\mathbb{Z}^2$-action defined by the shift on the compact group

$$X_{\bullet\bullet} = \{x \in \mathbb{F}_2^{\mathbb{Z}^2} \mid x_{\mathbf{n}+\mathbf{e}_1} + x_{\mathbf{n}+\mathbf{e}_2} + x_{\mathbf{n}} = 0 \text{ for all } \mathbf{n} \in \mathbb{Z}^2\},$$

where $\mathbb{F}_2 = \{0, 1\}$ denotes the field with two elements (see [38, Sect. 8.2]). As the notation suggests, having fixed the binary alphabet $\{0, 1\}$, this system is determined by its defining shape $\begin{smallmatrix}\bullet & \bullet\\ \bullet & \end{smallmatrix}$. In this section we will consider a simple instance of how the measurable structure of such a system varies as the defining shape is changed.

Example 5.5 (Reverse Ledrappier's Example). The reverse Ledrappier example is the $\mathbb{Z}^2$-action by shifts on the compact group

$$X_{\begin{smallmatrix}\bullet & \bullet\\ \bullet & \end{smallmatrix}} = \{x \in \mathbb{F}_2^{\mathbb{Z}^2} \mid x_{\mathbf{n}-\mathbf{e}_1} + x_{\mathbf{n}+\mathbf{e}_2} + x_{\mathbf{n}} = 0 \text{ for all } \mathbf{n} \in \mathbb{Z}^2\}.$$

Write $\mathsf{X}_{\begin{smallmatrix}\bullet & \bullet\\ \bullet & \end{smallmatrix}}$ for the measure-preserving $\mathbb{Z}^2$ system defined by the shift σ on $X_{\begin{smallmatrix}\bullet & \bullet\\ \bullet & \end{smallmatrix}}$, preserving Haar measure $m_{X_{\begin{smallmatrix}\bullet & \bullet\\ \bullet & \end{smallmatrix}}}$ defined on the Borel σ-algebra $\mathscr{B}_{X_{\begin{smallmatrix}\bullet & \bullet\\ \bullet & \end{smallmatrix}}}$, and similarly for the reverse shape $\begin{smallmatrix}\bullet & \bullet\\ & \bullet\end{smallmatrix}$. Notice that we write σ for the natural shift action on any group of the form $F^{\mathbb{Z}^2}$, where F denotes any finite group.

Recall from [38, Def. 6.7] that a *joining* of two measure-preserving systems $\mathsf{X} = (X, \mathscr{B}_X, \mu, T)$ and $\mathsf{Y} = (Y, \mathscr{B}_Y, \nu, S)$ is a Borel probability measure on $X \times Y$ that is invariant under $T \times S$, defined on $\mathscr{B}_X \otimes \mathscr{B}_Y$, and projects to the measures μ and ν on the X and Y coordinates respectively. This definition extends in a natural way to two measure-preserving actions of a group: the only change is that the joining measure is required to be invariant under the product group action. As in [38, Def. 6.14], we say that two group actions are disjoint if the only joining is the product of the two measures.

Theorem 5.6. *The systems* $\mathsf{X}_{\begin{smallmatrix}\bullet & \bullet\\ \bullet & \end{smallmatrix}}$ *and* $\mathsf{X}_{\begin{smallmatrix}\bullet & \bullet\\ & \bullet\end{smallmatrix}}$ *are disjoint.*

We will prove this by showing that one (and hence any) of the measure-preserving transformations on the group $X_{\begin{smallmatrix}\bullet & \bullet\\ \bullet & \end{smallmatrix}} \times X_{\begin{smallmatrix}\bullet & \bullet\\ & \bullet\end{smallmatrix}}$ defined by a joining has maximal entropy. Before embarking on the proof we assemble some properties of the two systems.

Lemma 5.7. *Write* $Y = \{x \in X_{\begin{smallmatrix}\bullet & \bullet\\ \bullet & \end{smallmatrix}} \mid x_{\mathbf{m}} = 0 \text{ for } m_1 \geqslant 0\}$. *Then the subgroup*

$$Z = \bigcup_{n=1}^{\infty} \sigma^{(-n,0)}(Y)$$

is dense in $X_{\begin{smallmatrix}\bullet & \bullet\\ \bullet & \end{smallmatrix}}$. *It follows that the Haar measure* $m_{X_{\begin{smallmatrix}\bullet & \bullet\\ \bullet & \end{smallmatrix}}}$ *is the only Borel probability measure on* $X_{\begin{smallmatrix}\bullet & \bullet\\ \bullet & \end{smallmatrix}}$ *invariant under translation by all elements of* Z, *and is the only* σ*-invariant probability measure invariant under translation by all elements of* Y.

Proof. It is enough to prove that Z is a dense subgroup of $X_{\begin{smallmatrix}\bullet & \bullet\\ \bullet & \end{smallmatrix}}$; the claim about invariant measures follows since a Borel measure is determined by how it integrates continuous functions (as in the proof that (3) $\implies$ (1) in [38, Th. 4.14]). Now $\sigma^{(0,\pm 1)}(Y) = Y$, so the subgroup Z is invariant under the whole shift σ. Thus in order to show that Z is dense, it is enough to show

that for any cylinder set C defined by specifying the coordinates in some square $\{(a_1, a_2) \mid 0 \leqslant a_1, a_2 \leqslant N\}$ contains an element of Z. Choose a sequence

$$(y_{(0,0)}, y_{(0,1)}, \ldots, y_{(0,2N)}) \tag{5.3}$$

(see Figure 5.1) chosen to ensure* that the only way to extend y to the coordinates in

$$\{(a_1, a_2) \mid 0 \leqslant a_1, a_2 \leqslant N\}$$

agrees with the condition defining the cylinder set C. We extend the finite sequence in equation (5.3) by 0s on the right and the left to define an element of $\{0,1\}^{\mathbb{Z}}$. Using these coordinates and the defining relation $\vdots$ we may define $y_{(n_1,n_2)}$ for $n_2 \geqslant 0$ uniquely. The coordinates of $y_{(n_1,n_2)}$ for $n_2 < 0$ are not determined by these choices. However, by choosing $y_{(2N,n_2)} = y_{(2N,0)}$ for all $n_2 \leqslant 0$ and applying the defining relation once more, this defines a point y in $\sigma^{(-2N-1,0)}Y \cap C$ as required.

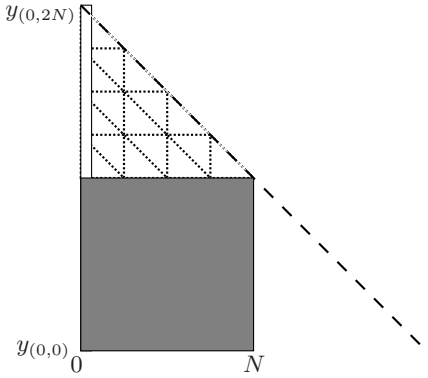

Fig. 5.1. The group Z meets each cylinder set.

□

We now turn to the main argument, which may be described as studying the *entropy geometry* of $X_{\vdots}$ and $X_{\vdots}$. Notice that for any finite alphabet group F and closed shift-invariant subgroup of $F^{\mathbb{Z}^2}$, the *state partition*

$$\xi = \left\{\{x \in F^{\mathbb{Z}^2} \mid x_{(0,0)} = f\} \mid f \in F\right\} \tag{5.4}$$

is a generator for the whole action in the sense that $\bigvee_{\mathbf{n}\in\mathbb{Z}^2} \sigma^{-\mathbf{n}}(\xi)$ is the Borel σ-algebra in $F^{\mathbb{Z}^2}$. Of course the state partition may or may not generate under the action of some subgroup $L \leqslant \mathbb{Z}^2$, depending on the exact

* By definition of $X_{\vdots}$: for example, $y_{(0,0)}$ and $y_{(0,1)}$ together determine the coordinate $y_{(1,0)} = y_{(0,0)} + y_{(0,1)}$, while $y_{(0,0)}, y_{(0,1)}, y_{(0,2)}$ together also determine $y_{(1,1)}$, and hence $y_{(2,0)} = y_{(1,0)} + y_{(1,1)}$.

rules defining the closed shift-invariant subgroup and its relation to the subgroup $L^{(34)}$. More generally, to any finite set of coordinates in $\mathbb{Z}^2$ there is a naturally associated partition defined by all the possible cylinder sets obtained by specifying those coordinates.

Proof of Theorem 5.6. Let $X = X_{\vdots} \times X_{\vdots} \subseteq \left(\mathbb{F}_2^2\right)^{\mathbb{Z}^2}$, write σ for the usual shift $\mathbb{Z}^2$-action on X, and let $\rho \in \mathsf{J}(\mathsf{X}_{\vdots}, \mathsf{X}_{\vdots})$ be a joining of the two systems (so ρ is a shift-invariant Borel probability measure on X that projects to Haar measure on each of the two coordinates). It will be helpful to think of $X_{\vdots}$ as being a 'top' layer and $X_{\vdots}$ as a 'bottom' layer. Notice that points in X obey the rule $\vdots$ in the top layer, obey the rule $\vdots$ in the bottom layer and, as may be checked directly, obey the rule

$$\begin{matrix} \cdot & \cdot \\ \cdot & \cdot \end{matrix} \tag{5.5}$$

in both layers at once. Alternatively, one can argue in terms of polynomials: the relation $\vdots$ corresponds* to annihilation by the polynomial $1 + u_1 + u_2$ in $\mathbb{F}_2[u_1^{\pm 1}, u_2^{\pm 1}]$ with support $\{(0,0),(1,0),(0,1)\}$, the relation $\vdots$ corresponds to annihilation by $1 + u_1^{-1} + u_2$, and the relation (5.5) corresponds to annihilation by the product

$$(1 + u_2 + u_2)(1 + u_1^{-1} + u_2) = u_1^{-1} + u_1 + u_1u_2 + u_1^{-1}u_2 + u_2^2,$$

whose support $\{(-1,0),(1,0),(-1,1),(1,1),(0,2)\}$ is illustrated in (5.5).

By Proposition 1.16 we have

$$h_\rho(\sigma^{(1,0)}) = \sup_\xi H_\rho\left(\xi \middle| \bigvee_{i=1}^{\infty} \sigma^{-(i,0)}(\xi) \right),$$

where the supremum is taken over all finite partitions. As usual (by Theorem 1.20 or 4.14) the supremum is attained by a partition that generates under $\sigma^{(1,0)}$ or by taking a limit along a sequence of partitions that generate. Define ξ_ℓ to be the partition defined by specifying the $(4\ell + 1)$ coordinates illustrated in Figure 5.2.

The sequence of partitions (ξ_ℓ) satisfies the hypothesis of Theorem 4.14, so

$$h_\rho(\sigma^{(1,0)}) = \lim_{\ell\to\infty} H_\rho\left(\xi_\ell \middle| \bigvee_{i=1}^{\infty} \sigma^{-(i,0)}(\xi_\ell) \right). \tag{5.6}$$

Geometrically, $H_\rho\left(\xi_\ell \middle| \bigvee_{i=1}^{\infty} \sigma^{-(i,0)}(\xi_\ell) \right)$ is the entropy corresponding to learning the shaded coordinates given complete knowledge of the coordinates to the right, as illustrated in Figure 5.3. In the top layer $X_{\vdots}$ the coordinates at $(1,1),\ldots,(2\ell,2\ell)$ are determined, while in the bottom layer $X_{\vdots}$ the coordinates at $(0,-1),\ldots,(0,-2\ell)$ are determined. The remaining coordinates are

* Explanation somehow – but avoiding duality? Need to say something.

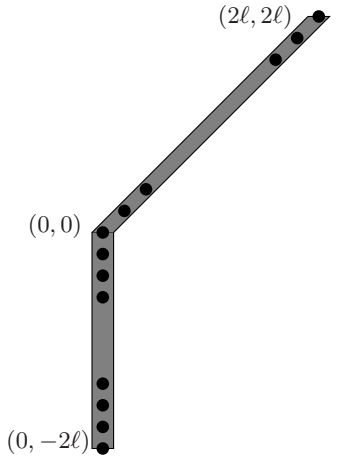

Fig. 5.2. The coordinates defining the partition ξ_ℓ.

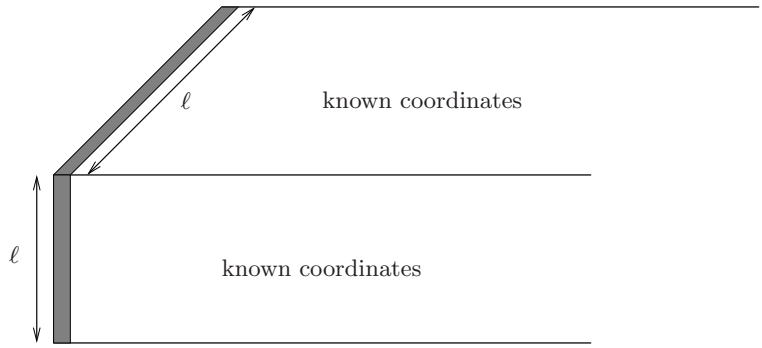

Fig. 5.3. The entropy calculation $H_\rho(\xi_\ell | \bigvee_{i=1}^{\infty} \sigma^{-(i,0)}(\xi_\ell))$.

determined once a single choice is made in the top layer and a single choice is made in the bottom layer at one (arbitrary) undetermined coordinate. For example, we may choose the top layer coordinate at $(0, -\ell)$ marked $\Diamond$ and the bottom layer coordinate at (ℓ, ℓ) marked $\blacklozenge$ as in Figure 5.4.

Once these choices have been made, the defining relation $\because$ in the top layer propagates the choice to all of the top layer shaded coordinates, and the defining relation $\because$ in the bottom layer propagates the choice to all of the bottom layer shaded coordinates. Thus

$$H_\rho\left(\xi_\ell \Big| \bigvee_{i=1}^{\infty} \sigma^{-(i,0)}\xi_\ell\right) = H_\rho\left(\sigma^{(0,-\ell)}\xi \vee \sigma^{(-\ell,-\ell)}\xi \Big| \bigvee_{i=1}^{\infty} \sigma^{-(i,0)}\xi_\ell\right) \qquad (5.7)$$

where, as in equation (5.4), we write ξ for the state partition. Having made these choices, all other coordinates are determined. In particular, equation (5.6) shows that

$$h_\rho(\sigma^{(1,0)}) \leqslant \log 4.$$

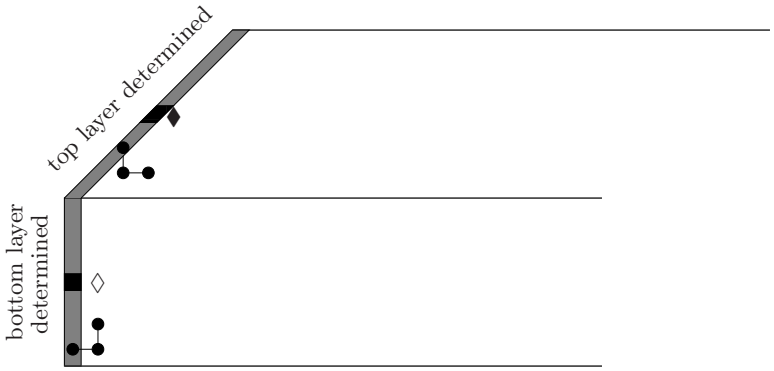

Fig. 5.4. The choice required to determine the shaded region.

We now show how this combinatorial argument can be used to give a geometrical decomposition of the entropy. Write

$$\mathsf{H}_1 = \{\mathbf{m} \in \mathbb{Z}^2 \mid \mathbf{m} \cdot (-1, 0) < 0\}$$

for the right half-plane in $\mathbb{Z}^2$, and

$$\mathsf{H}_2 = \{\mathbf{m} \in \mathbb{Z}^2 \mid \mathbf{m} \cdot (-1, 1) < 0\},$$

for the lower right diagonal half-plane in $\mathbb{Z}^2$. Notice that $\mathsf{H}_1 \cap \mathsf{H}_2$ is approximated by the region of known coordinates in Figure 5.3 for large ℓ. Write $\mathscr{C}_\ell = \bigvee_{i=1}^{\infty} \sigma^{(-i,0)}\xi_\ell$ and $\mathscr{C} = \bigvee_{\ell \geqslant 1} \mathscr{C}_\ell$. By Proposition 1.7 and equation (5.7),

$$\begin{aligned} H_\rho\left(\xi_\ell \Big| \bigvee_{i=1}^{\infty} \sigma^{-(i,0)}\xi_\ell\right) &= H_\rho\left(\sigma^{(0,-\ell)}\xi \big| \mathscr{C}_\ell\right) + H_\rho\left(\sigma^{-(\ell,\ell)}\xi \big| \mathscr{C}_\ell \vee \sigma^{(0,-\ell)}\xi\right) \\ &= H_\rho\left(\xi \big| \sigma^{(0,\ell)}\mathscr{C}_\ell\right) + H_\rho\left(\xi \big| \sigma^{(\ell,\ell)}\mathscr{C}_\ell \vee \sigma^{(\ell,0)}\xi\right) \quad (5.8) \end{aligned}$$

for any $\ell \geqslant 1$. Now

$$\sigma^{(0,\ell)}\mathscr{C}_\ell \subseteq \bigvee_{\mathbf{n} \in \mathsf{H}_1} \sigma^{-\mathbf{n}}\xi_\ell$$

and

$$\sigma^{(\ell,\ell)}\mathscr{C}_\ell \vee \sigma^{(\ell,0)}\xi \subseteq \bigvee_{\mathbf{n} \in \mathsf{H}_2} \sigma^{-\mathbf{n}}\xi_\ell$$

so, by Proposition 1.7 and equation (5.8),

$$H_\rho\left(\xi_\ell\big|\bigvee_{i=1}^{\infty}\sigma^{-(i,0)}\xi_\ell\right)\geqslant H_\rho\left(\xi\big|\bigvee_{\mathbf{n}\in\mathsf{H}_1}\sigma^{-\mathbf{n}}\mathscr{C}_\ell\right)+H_\rho\left(\xi\big|\bigvee_{\mathbf{n}\in\mathsf{H}_2}\sigma^{-\mathbf{n}}\mathscr{C}_\ell\right).$$

By equation (5.6) we deduce that

$$h_\rho(\sigma^{(1,0)})\geqslant H_\rho\left(\xi\big|\bigvee_{\mathbf{n}\in\mathsf{H}_1}\sigma^{-\mathbf{n}}\mathscr{C}\right)+H_\rho\left(\xi\big|\bigvee_{\mathbf{n}\in\mathsf{H}_2}\sigma^{-\mathbf{n}}\mathscr{C}\right).\tag{5.9}$$

On the other hand, the increasing martingale theorem (see [38, Th. 5.5]) implies that

$$H_\rho\left(\xi\big|\sigma^{(0,\ell)}\mathscr{C}_\ell\right)\to H_\rho\left(\xi\big|\bigvee_{\mathbf{n}\in\mathsf{H}_1}\sigma^{-\mathbf{n}}\mathscr{C}\right)$$

since $\sigma^{(0,\ell)}\mathscr{C}_\ell\nearrow\bigvee_{\mathbf{n}\in\mathsf{H}_1}\sigma^{-\mathbf{n}}\xi_\ell$, and

$$H_\rho\left(\xi\big|\sigma^{(\ell,\ell)}\mathscr{C}_\ell\vee\sigma^{(\ell,0)}\xi\right)\to H_\rho\left(\xi\big|\bigvee_{\mathbf{n}\in\mathsf{H}_2}\sigma^{-\mathbf{n}}\mathscr{C}\right)$$

similarly, so equation (5.6) and equation (5.8) give

$$h_\rho(\sigma^{(1,0)})\leqslant H_\rho\left(\xi\big|\bigvee_{\mathbf{n}\in\mathsf{H}_1}\sigma^{-\mathbf{n}}\mathscr{C}\right)+H_\rho\left(\xi\big|\bigvee_{\mathbf{n}\in\mathsf{H}_2}\sigma^{-\mathbf{n}}\mathscr{C}\right).$$

Together with equation (5.9), this gives

$$h_\rho(\sigma^{(1,0)})=H_\rho\left(\xi\big|\bigvee_{\mathbf{n}\in\mathsf{H}_1}\sigma^{-\mathbf{n}}\mathscr{C}\right)+H_\rho\left(\xi\big|\bigvee_{\mathbf{n}\in\mathsf{H}_2}\sigma^{-\mathbf{n}}\mathscr{C}\right),\tag{5.10}$$

in which the first term is no larger than $\log 2$ by the argument concerning Figure 5.4.

Now

$$\bigvee_{\mathbf{n}\in\mathsf{H}_2}\sigma^{-\mathbf{n}}\mathscr{C}\supseteq\mathscr{B}_1=\mathscr{B}_{X_{\bullet\bullet}}\times\mathscr{N}_{X_{\bullet\bullet}}\tag{5.11}$$

since in the top layer (a copy of $\mathsf{X}_{\bullet\bullet}$) the coordinates in H_2 determine everything. On the other hand, $\mathscr{B}_1$ is the σ-algebra corresponding to the factor $\mathsf{X}_{\bullet\bullet}$ of the dynamical system

$$(X,\mathscr{B}_{X_{\bullet\bullet}}\otimes\mathscr{B}_{X_{\bullet\bullet}},\rho,\sigma^{(1,0)}),$$

so by Corollary 4.15 we can also decompose the entropy as

$$h_\rho(\sigma^{(1,0)})=h_{m_{\bullet\bullet}}(\sigma^{(1,0)})+h_\rho(\sigma^{(1,0)}\big|\mathscr{B}_1)\tag{5.12}$$

where we write $m_{\bullet\bullet}, m_{\bullet\,\bullet}$ for the Haar measures on $X_{\bullet\bullet}, X_{\bullet\,\bullet}$ respectively. Going through the arguments in equations (5.6) to (5.10) again, but for $h_\rho(\sigma^{(1,0)}|\mathscr{B}_1)$ and conditioning throughout on the σ-algebra $\mathscr{B}_1$, we obtain

$$h_\rho(\sigma^{(1,0)}\big|\mathscr{B}_1) = H_\rho\left(\xi\big|\bigvee_{\mathbf{n}\in\mathsf{H}_1}\sigma^{-\mathbf{n}}\mathscr{C}\vee\mathscr{B}_1\right) + H_\rho\left(\xi\big|\bigvee_{\mathbf{n}\in\mathsf{H}_2}\sigma^{-\mathbf{n}}\xi\vee\mathscr{B}_1\right).$$

However,

$$\bigvee_{\mathbf{n}\in\mathsf{H}_1}\sigma^{-\mathbf{n}}\mathscr{C}\vee\mathscr{B}_1 = \mathscr{B}_{X_{\bullet\bullet}}\otimes\mathscr{B}_{X_{\bullet\,\bullet}}$$

since the top layer $\mathsf{X}_{\bullet\bullet}$ corresponds to $\mathscr{B}_1$ and for the bottom layer $\mathsf{X}_{\bullet\,\bullet}$ all coordinates are determined by the coordinates in H_1.

Therefore, by equation (5.11),

$$h_\rho(\sigma^{(1,0)}\big|\mathscr{B}_1) = H_\rho\left(\xi\big|\bigvee_{\mathbf{n}\in\mathsf{H}_2}\sigma^{-\mathbf{n}}(\xi)\right).$$

Together with equation (5.10) and equation (5.12) we therefore have

$$\log 2 = h_{m_{\bullet\bullet}}(\sigma^{(1,0)}) = H_\rho\left(\xi\big|\bigvee_{\mathbf{n}\in\mathsf{H}_1}\sigma^{-\mathbf{n}}(\xi)\right).$$

By the same argument, for any $N\geqslant 1$,

$$N\log 2 = H_\rho\left(\bigvee_{n=0}^{N-1}\sigma^{(n,0)}(\xi)\big|\bigvee_{\mathbf{n}\in\mathsf{H}_1}\sigma^{-\mathbf{n}}(\xi)\right)$$

while $\bigvee_{n=0}^{N-1}\sigma^{-(n,0)}(\xi)$ contains no more than 2^N elements. It follows by Proposition 1.7(1) and Proposition 1.5 that each atom of $\bigvee_{n=0}^{N-1}\sigma^{-(n,0)}(\xi)$ must have ρ-measure $\frac{1}{2^N}$. Thus ρ is invariant under rotation by any element of the group $Y\subseteq X_{\bullet\,\bullet}\times\{0\}$, so by Lemma 5.7 it is invariant under rotation by $X_{\bullet\bullet}\times\{0\}$. This forces ρ to be $m_{\bullet\bullet}\times\nu$ for some measure ν on $X_{\bullet\,\bullet}$, and ν must be $m_{\bullet\,\bullet}$ since it is the projection of ρ onto the bottom layer. This proves Theorem 5.6. □

Exercises for Section 5.2

Exercise 5.2.1. (35) Show that the system $\mathsf{X}_{\bullet\bullet}$ has many invariant measures, as follows.
(a) Find an invariant probability measure other than $m_{\bullet\bullet}$ for which $\sigma^{(1,0)}$ has positive entropy.
(b) Find a non-trivial invariant closed subset in $\mathsf{X}_{\bullet\bullet}$.

Exercise 5.2.2. (36) Show that $X_{\bullet\bullet}$ does exhibit a certain weak form of measure rigidity by showing that if ρ is an invariant Borel probability measure on $X_{\bullet\bullet}$ and $\sigma^{(1,0)}$ is mixing with respect to ρ, then $\rho = m_{\bullet\bullet}$.

5.3 Rigidity of Positive Entropy Measures for $\times 2, \times 3$: Rudolph's Theorem

It is clear that the maps $S_2 : x \mapsto 2x$ modulo 1 and $S_3 : x \mapsto 3x$ modulo 1 on the circle $\mathbb{T}$ each have many invariant probability measures and many closed invariant sets, both infinite and finite. Furstenberg [46] showed that any closed invariant set that is simultaneously invariant under both maps is either the whole circle or finite, and asked if the analog of this result for probability measures held true. That is, if an ergodic probability measure is invariant under both maps, can it be anything other than Lebesgue measure or an atomic measure? A measure invariant under both maps will be called a $\times 2, \times 3$-invariant measure, and the result sought says that under suitable hypotheses the only $\times 2, \times 3$-invariant probability measure is Lebesgue measure.

Progress was made on this question by Lyons [91], who proved the result for measures with completely positive entropy with respect to one of the maps. The assumption of completely positive entropy was reduced to just positive entropy by Rudolph [126], and this is the result we present here. Johnson [66] later removed the coprimality assumption, extending(37) Rudolph's theorem to measures invariant under multiplication by 5 and 10 for example.

Furstenberg's topological result described in Section 5.1 motivated the question about what a non-atomic probability measure invariant under both of the maps $\times 2$ and $\times 3$ might look like. The central problem remains open, but under an additional entropy condition the problem is now well understood by work of Rudolph [126].

Theorem 5.8 (Rudolph). *Let S_2 denote $x \mapsto 2x \pmod 1$ and S_3 denote $x \mapsto 3x \pmod 1$ on $\mathbb{T}$. Let μ be an S_2, S_3-invariant probability measure on $\mathbb{T}$ which is ergodic with the property that $h_\mu(S_2^m S_3^n) > 0$ for some $m, n \in \mathbb{N}$. Then μ is the Lebesgue measure $m_\mathbb{T}$ on $\mathbb{T}$.*

Here ergodicity is meant with respect to the joint action, that is if $A \subseteq \mathbb{T}$ is measurable and $S_2^{-1}A = S_3^{-1}A = A \pmod \mu$ then $\mu(A) = 0$ or 1.

In order to prove this, it is convenient to work with the invertible extension of the $\mathbb{N}^2$-action in which the pair (m, n) is sent to the map $S_2^m S_3^n$. To make the invertible extension, define

$$X = \{x \in \mathbb{T}^{\mathbb{Z}^2} \mid x_{\mathbf{n}+\mathbf{e}_1} = 2x_\mathbf{n} \text{ and } x_{\mathbf{n}+\mathbf{e}_2} = 3x_\mathbf{n} \text{ for all } \mathbf{n} \in \mathbb{Z}^2\}.$$

The conditions defining X as a subset of $\mathbb{T}^{\mathbb{Z}^2}$ are closed and homogeneous, so X is a compact abelian group. The group X is invariant under the shift action of $\mathbb{Z}^2$ defined by

$$(x_\mathbf{n})_\mathbf{n} \overset{\text{action of } \mathbf{m}}{\longrightarrow} (x_{\mathbf{n}+\mathbf{m}})_\mathbf{n}.$$

Write T_2 for the left shift by $\mathbf{e}_1$ and T_3 for the down shift by $\mathbf{e}_2$. Thus

$$(T_2^m T_3^n x)_{(a,b)} = x_{(a+m,b+n)}$$

for all $a, b, m, n \in \mathbb{Z}$, so T_2, T_3 are commuting automorphisms of X. Together they define a $\mathbb{Z}^2$-action that is the invertible extension of the $\mathbb{N}^2$-action defined by the commuting endomorphisms S_2 and S_3.

Write $\pi_{\mathbf{0}} : X \to \mathbb{T}$ for the surjective homomorphism defined by

$$\pi_{\mathbf{0}}(x) = x_{(0,0)}.$$

By the definition of X, for any $m, n \geqslant 0$ the diagram

$$\begin{array}{ccc} X & \xrightarrow{T_2^m T_3^n} & X \\ {\scriptstyle \pi_{\mathbf{0}}}\downarrow & & \downarrow{\scriptstyle \pi_{\mathbf{0}}} \\ \mathbb{T} & \xrightarrow[S_2^m S_3^n]{} & \mathbb{T} \end{array}$$

commutes. Every T_2, T_3-invariant measure μ_X on X defines a S_2, S_3-invariant measure $\mu_{\mathbb{T}} = (\pi_{\mathbf{0}})_* \mu_X$ on $\mathbb{T}$ and *vice versa*. Moreover, ergodicity is preserved: μ_X is ergodic if and only if $\mu_{\mathbb{T}}$ is.

It will be useful to visualize the group X as a lattice of circles in which a step to the right is an application of S_2, a step up is an application of S_3 and therefore each point in $\mathbb{T}$ at a given lattice site has two choices for the point to its left and three choices for the point below it. This is illustrated in Figure 5.5.

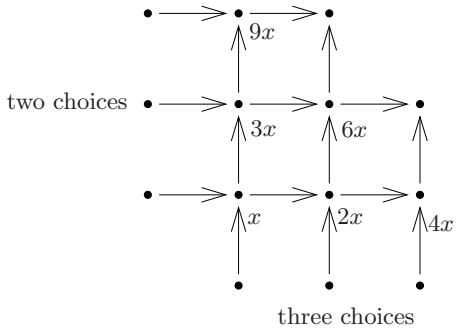

Fig. 5.5. The invertible extension of $\times 2, \times 3$.

Define a partition on $\mathbb{T}$ by

$$\xi_{\mathbb{T}} = \{[0, \tfrac{1}{6}), [\tfrac{1}{6}, \tfrac{2}{6}), \ldots, [\tfrac{5}{6}, 1)\}$$

and induce a partition on X by

$$\xi_X = \pi_{\mathbf{0}}^{-1} \xi_{\mathbb{T}}.$$

Lemma 5.9. *The partition $\xi_{\mathbb{T}}$ is a generator for S_2 and is a generator for S_3. In general the partition ξ_X is not a generator for T_2, nor is it a generator for T_3. Nonetheless, for any invariant measure μ_X on X,*

$$h_{\mu_X}(T_2, \xi_X) = h_{\mu_X}(T_2) = h_{\mu_{\mathbb{T}}}(S_2) = h_{\mu_{\mathbb{T}}}(S_2, \xi_{\mathbb{T}}).$$

Proof. The only non-trivial part of this is to show that

$$h_{\mu_X}(T_2, \xi_X) = h_{\mu_X}(T_2).$$

Let $\mathscr{C} = \sigma\left(\bigvee_{i=-\infty}^{\infty} T_2^{-i}\xi_X\right)$. In Figure 5.5, $\mathscr{C}$ is the σ-algebra generated by the coordinates in the upper half-plane. Figure 5.6 shows what $[x]_{\mathscr{C}}$ tells us about x.

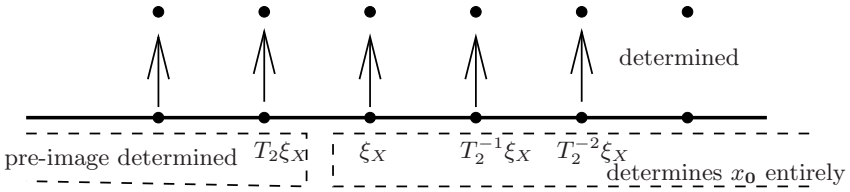

Fig. 5.6. How much is determined about x if $[x]_{\mathscr{C}}$ is known.

Now $T_3\xi_X$ (and the associated factor $T_3\mathscr{C}$) gives the same entropy to T_2. Since $T_3^n\mathscr{C} \nearrow \mathscr{B}_X$, Theorem 4.14 shows that

$$h_{\mu_X}(T_2) = \lim_{n\to\infty} h_{\mu_X}(T_2|_{T_3^n\mathscr{C}}) = h_{\mu_X}(T_2, \xi_X).$$

□

The key step in the proof is a relationship between the entropy of T_2 and T_3 which holds for the $\times 2, \times 3$-invariant measure $\mu = \mu_X$ on X.

Proposition 5.10. *For any σ-algebra $\mathscr{S}$ invariant under T_2 and T_3,*

$$h_\mu(T_3|\mathscr{S}) = \frac{\log 3}{\log 2} h_\mu(T_2|\mathscr{S}).$$

In particular,

$$h_\mu(T_3) = \frac{\log 3}{\log 2} h_\mu(T_2).$$

Proof. Let $\kappa = \frac{\log 3}{\log 2}$. The partition $\bigvee_{i=0}^{n-1} S_3^{-i}\xi_{\mathbb{T}}$ comprises intervals of length $\frac{1}{2\cdot 3^n}$, while the partition $\bigvee_{i=0}^{m-1} S_2^{-i}\xi_{\mathbb{T}}$ comprises intervals of length $\frac{1}{2^m\cdot 3}$. Write $\ell_3(n) = \frac{1}{2\cdot 3^n}$ and $\ell_2(m) = \frac{1}{2^m\cdot 3}$. If $m = \lfloor \kappa n \rfloor$ then

$$\frac{1}{2}\ell_2(m) = \frac{1}{2^{m+1}\cdot 3} \leqslant \frac{1}{3\cdot 3^n} \leqslant \ell_3(n) \leqslant \frac{1}{2\cdot 2^m} = \frac{3}{2}\ell_2(m),$$

so each interval in $\bigvee_{i=0}^{n-1} S_3^{-i}\xi_{\mathbb{T}}$ intersects at most three intervals in $\bigvee_{i=0}^{m-1} S_2^{-i}\xi_{\mathbb{T}}$ and *vice versa*. It follows that

$$H_\mu\left(\bigvee_{i=0}^{n-1} T_3^{-i}\xi_X \Big| \bigvee_{j=0}^{m-1} T_2^{-j}\xi_X\right) \leqslant \log 3,$$

and

$$H_\mu\left(\bigvee_{j=0}^{m-1} T_2^{-j}\xi_X \Big| \bigvee_{i=0}^{n-1} T_3^{-i}\xi_X\right) \leqslant \log 3,$$

so

$$\begin{aligned} H_\mu\left(\bigvee_{i=0}^{n-1} T_3^{-i}\xi_X \Big| \mathscr{S}\right) &\leqslant H_\mu\left(\bigvee_{i=0}^{n-1} T_3^{-i}\xi_X \vee \bigvee_{j=0}^{m-1} T_2^{-j}\xi_X \Big| \mathscr{S}\right) \\ &\leqslant H_\mu\left(\bigvee_{j=0}^{m-1} T_2^{-j}\xi_X \Big| \mathscr{S}\right) + \underbrace{H_\mu\left(\bigvee_{i=0}^{n-1} T_3^{-i}\xi_X \Big| \bigvee_{j=0}^{m-1} T_2^{-j}\xi_X\right)}_{\leqslant \log 3}. \end{aligned}$$

Now let $n \to \infty$ (with $m = \lfloor \kappa n \rfloor$) to deduce that

$$h_\mu(T_3 \big| \mathscr{S}) \leqslant \underbrace{\lim_{n\to\infty} \frac{m}{n}}_{\to\kappa} h_\mu(T_2 \big| \mathscr{S}) + 0.$$

The same argument with the role of S_2 and S_3 reversed gives the reverse inequality. □

A similar argument applies to any map $T_2^m T_3^n$ with $m \geqslant 0$ and $n \geqslant 0$, so we deduce that the following properties of a measure μ are equivalent:

- $h_\mu(T_2) > 0$;
- $h_\mu(T_3) > 0$;
- $h_\mu(T_2^m T_3^n) > 0$ for some $m \geqslant 0$, $n \geqslant 0$;
- $h_\mu(T_2^m T_3^n) > 0$ for all $(m,n) \neq (0,0)$ with $m \geqslant 0$, $n \geqslant 0$.

Corollary 5.11. *The maps T_2 and T_3 of (X,μ) have the same Pinsker algebra.*

Proof. Let $\mathscr{P}(T_2)$ be the Pinsker algebra of T_2. Then it is easy to check that $T_3\mathscr{P}(T_2) = \mathscr{P}(T_2)$. Proposition 5.10 then implies that

$$\begin{aligned} h_\mu\left(T_3\big|\mathscr{P}(T_2)\right) &= \kappa h_\mu\left(T_2\big|\mathscr{P}(T_2)\right) \\ &= \kappa h_\mu(T_2) \\ &= h_\mu(T_3). \end{aligned}$$

This in turn gives $h_\mu(T_3\big|_{\mathscr{P}(T_2)}) = 0$ by Corollary 4.15, so $\mathscr{P}(T_2) \subseteq \mathscr{P}(T_3)$. The corollary follows by symmetry of the argument. □

Proving Theorem 5.8 requires a more geometric understanding of entropy. Let

$$\mathscr{A}_1 = \sigma\left(\bigvee_{i=1}^{\infty} T_2^{-i}\xi_X\right) = T_2^{-1}\pi_{\mathbf{0}}^{-1}\mathscr{B}_{\mathbb{T}},$$

illustrated in Figure 5.7.

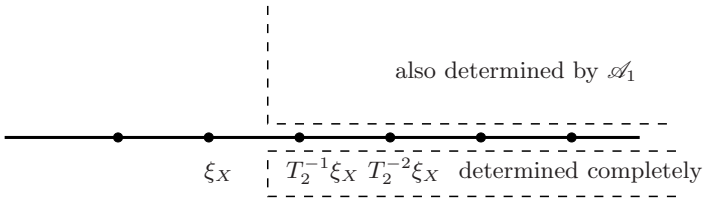

Fig. 5.7. The σ-algebra $\mathscr{A}_1$.

Knowledge of $S_2(y)$ and $S_3(y)$ in $\mathbb{T}$ determines $y \in \mathbb{T}$ uniquely, so

$$\xi_X \vee T_3^n\mathscr{A}_1 = T_3^n\xi_X \vee T_3^n\mathscr{A}_1 \tag{5.13}$$

(cf. Figure 5.8).

Fig. 5.8. Proof that $\xi_X \vee T_3^n\mathscr{A}_1 = T_3^n\xi_X \vee T_3^n\mathscr{A}_1$.

Lemma 5.9 shows that

$$\begin{aligned} h_\mu(T_2) &= H_\mu(\xi_X\big|\mathscr{A}_1) \\ &= H_\mu(T_3^n\xi_X\big|T_3^n\mathscr{A}_1) \quad \text{(since } \mu \text{ is invariant under } T_3\text{)} \\ &= H_\mu(\xi_X\big|T_3^n\mathscr{A}_1) \quad \text{(by equation (5.13)).} \end{aligned}$$

It follows that if

$$\mathscr{A} = \bigvee_{i=1}^{\infty} \bigvee_{j=-\infty}^{\infty} T_2^{-i} T_3^{-j} \xi_X,$$

the σ-algebra determined by the coordinates in the half-plane

$$\{(m,n) \in \mathbb{Z}^2 \mid m > 0\},$$

then by Proposition 4.8,

$$h_\mu(T_2) = H_\mu(\xi_X \big| \mathscr{A}).$$

The next step is to identify the $\mathscr{A}$-atoms more closely.

Lemma 5.12. *$G = [0]_{\mathscr{A}}$ is a closed subgroup of X. Moreover, $[x]_{\mathscr{A}} = G + x$ for $x \in X$.*

Proof. The projection onto the coordinates in the half-space

$$\{(m,n) \in \mathbb{Z}^2 \mid m > 0\}$$

is a continuous homomorphism, so its kernel is a closed subgroup of X, and the pre-images of other points are the cosets of this kernel. □

Since the atom $[x]_{\mathscr{A}}$ supports the measure $\mu_x^{\mathscr{A}}$ and $[x]_{\mathscr{A}}$ is a coset of G, we can define for almost every $x \in X$ a measure ν_x on G by $\nu_x(A) = \mu_x^{\mathscr{A}}(A + x)$.

Lemma 5.13. *$G \cong \varprojlim \mathbb{Z}/2^n\mathbb{Z} = \mathbb{Z}_2$, the 2-adic integers.*

The isomorphism between G and the projective limit $\mathbb{Z}_2$ may be described as follows. A 2-adic integer

$$a = \sum_{n=0}^{\infty} a_n 2^n$$

with $a_n \in \{0,1\}$ for all $n \geqslant 0$, determines a point $g \in G \subseteq X$ with

$$g_{(-n,0)} = \sum_{m=0}^{n} a_m 2^{m-n-1} = \sum_{m=0}^{\infty} a_m 2^{m-n-1} \pmod 1.$$

Proof of Lemma 5.13. Define a map $\widetilde{\pi}_n : G \to \mathbb{T}$ by $x \mapsto x_{(-n,0)} \in \frac{1}{2^{n+1}}\mathbb{Z}/\mathbb{Z}$, and write $\pi_n : G \to \mathbb{Z}/2^{n+1}\mathbb{Z}$ for the map $\pi_n = 2^{n+1}\widetilde{\pi}_n$. Since

$$2x_{(-n-1,0)} = x_{(-n,0)}$$

we have

$$2\widetilde{\pi}_{n+1} = \widetilde{\pi}_n$$

and $\pi_{n+1} = \pi_n \pmod{2^n}$ for all $n \geqslant 1$. This defines the isomorphism between G and the projective limit. □

Proposition 5.14. *The map $x \mapsto \nu_x$ is Pinsker (measurable with respect to the Pinsker algebra) with respect to T_2 and with respect to T_3.*

PROOF. Fix some $n \geqslant 1$. It is enough to show that $g_k(x) = \nu_x(\pi_n^{-1}k)$ is Pinsker for any k in $\mathbb{Z}/2^{n+1}\mathbb{Z}$. We claim that g_k is in fact periodic with respect to T_3. For this, note that multiplication by 3 is invertible modulo 2^{n+1} and recall that $T_3^{-1}\mathscr{A} = \mathscr{A}$. Now

$$\begin{aligned} g_k(T_3x) &= \mu^{\mathscr{A}}_{T_3x}\left(T_3x + \pi_n^{-1}k\right) \\ &= \mu_x^{T_3^{-1}\mathscr{A}}\left(x + T_3^{-1}\pi_n^{-1}k\right) \\ &= \mu_x^{\mathscr{A}}\left(x + \pi_n^{-1}(k/3)\right) = g_{k/3}(x). \end{aligned}$$

For any n there is some m with $3^m \equiv 1$ modulo 2^{n+1} (here we are really using the fact that $\gcd(2,3) = 1$; for natural numbers with a common factor a different argument is needed). For this value of m,

$$g_k(T_3^m x) = g_{3^{-m}k}(x) = g_k(x)$$

as claimed. By Corollary 5.11, ν_x is also Pinsker with respect to T_2. □

Proposition 5.15. *For almost every $x \in X$, the measure ν_x is the Haar measure m_{G_x} of a closed subgroup $G_x \leqslant G$.*

PROOF. Since $T_3^k \bigvee_{i=-\infty}^{\infty} T_2^{-i}\xi_X \nearrow \mathscr{B}_X$ as $k \to \infty$,

$$\mathscr{S} = \bigvee_k T_3^k \bigcap_{n=0}^{\infty} \bigvee_{i=n}^{\infty} T_2^{-i}(\xi_X) \subseteq \mathscr{A},$$

so ν_x is $\mathscr{A}$-measurable. Certainly $\mu_x^{\mathscr{A}}$ is $\mathscr{A}$-measurable (by [38, Th. 5.14]). Choose a null set N with the property that for $x, y \in [x]_{\mathscr{A}} \setminus N$,

$$\mu_x^{\mathscr{A}} = \mu_y^{\mathscr{A}},$$

$$\mu_x^{\mathscr{A}} - x = \nu_x = \nu_y = \mu_x^{\mathscr{A}} - y = \nu_x - x + y,$$

and

$$\mu_z^{\mathscr{A}}(N) = 0$$

almost everywhere. Note that this shows $\nu_x(N - x) = 0$, and so for ν_x-almost every $g \in G$ we have $y = x + g \in [x]_{\mathscr{A}} \setminus N$ and so

$$\nu_x = \nu_x + g$$

for ν_x-almost every $g \in G$. This implies that $G_x = \mathrm{Supp}(\nu_x)$ is a closed subgroup of G. □

PROOF OF THEOREM 5.8. By assumption, $h_\mu(T_2) > 0$, so

$$I_\mu(\xi_X \,|\, \mathscr{A}) > 0$$

on a set B of positive measure. On the other hand

$$\begin{aligned} I_\mu(\xi_X \,|\, \mathscr{A}) &= -\log \mu_x^{\mathscr{A}}(P(x)) \qquad (\text{for } x \in P(x) \in \mathscr{A}) \\ &= -\log \nu_x(\pi_{\mathbf{0}}^{-1}0) \\ &= -\log g_0(x), \end{aligned}$$

so $-\log g_0(x) > 0$. Thus G_x contains an element

$$g = \sum_{i=0}^{\infty} g_i 2^i$$

with $g_0 = 1$. However, G is topologically generated by any such element (since $\mathbb{Z}/2^{n+1}\mathbb{Z}$ is generated by any odd number). Thus $G_x = G$ for almost every $x \in B$. Since $g_0(x) = g_0(T_3 x)$ we know that the set B is T_3-invariant.

We claim that $G_x = G$ for almost every x. To prove this, it is enough to show that $B = X$.

For almost every $x \in B$, $\mu_x^{\mathscr{A}} = x + m_G$ (this property characterizes membership of B). Now

$$\mu_x^{T_2\mathscr{A}} = \mu_x^{\mathscr{A} \vee \xi_X} = \left(\mu_x^{\mathscr{A}}\right)_x^{\xi_X} = x + m_{T_2 G}.$$

Since

$$\mu^{\mathscr{A}}_{T_2^{-1}x} = \left(T_2^{-1}\right)_* \mu_x^{T_2\mathscr{A}} = T_2^{-1}x + m_G,$$

we have $T_2^{-1}x \in B$. Thus B is T_2-invariant. By the assumption of ergodicity, this implies that $B \underset{\mu}{=} X$. It follows that μ is invariant under translation by elements of G, since for any measurable set C and $g \in G$,

$$\begin{aligned} \mu(C+g) &= \int \mu_x^{\mathscr{A}}(C+g)\,\mathrm{d}\mu(x) \\ &= \int \mu_x^{\mathscr{A}}(C)\,\mathrm{d}\mu(x) \\ &= \mu(C). \end{aligned}$$

This invariance extends: for any $n \geqslant 1$,

$$\mu(C + T_2^{-n}g) = \mu(T_2^{-n}(T_2^n C + g)) = \mu(T_2^n C) = \mu(C),$$

so μ is invariant under translation by $\bigcup_{n=0}^{\infty} T_2^{-n}G$. The image under $\pi_{\mathbf{0}}$ of $\bigcup_{n=0}^{\infty} T_2^{-n}G$ is the dense subgroup $\{\frac{k}{2^{n+1}} \mid k \in \mathbb{Z}, n \geqslant 0\} \leqslant \mathbb{T}$; it follows that $\mu_{\mathbb{T}} = (\pi_{\mathbf{0}})_*\mu$ is Lebesgue measure on $\mathbb{T}$. Hence $\mu = m_X$ is Haar measure on X. □

Exercises for Section 5.3

Exercise 5.3.1. Extend Example 5.2 to show that a lacunary semigroup has many closed invariant subsets for its natural action on the circle by the following steps.

(1) For any $k > 1$, finite set $F \subseteq \mathbb{N}$, and set $A \subseteq \{0, \ldots, k-1\}^F$, show that the set

$$\left\{ x \in \mathbb{T} \mid x = \sum_{n=1}^{\infty} e_n k^{-n}, e|_{F+n} \in A \text{ for all } n \geqslant 1 \right\}$$

(where $e|_{F+n}$ means the projection of $e = (e_n)$ onto the set of coordinates $F + n \subseteq \mathbb{N}$) is invariant under S_k.

(2) More generally, show that for any $k > 1$ the map S_k has uncountably many closed invariant sets.

Exercise 5.3.2. [(38)] Let μ be an S_3-invariant and ergodic probability measure on $\mathbb{T}$.
(a) Prove that and weak*-limit of $\frac{1}{n}\sum_{j=0}^{n-1}(S_2)_*^j\mu$ is of the form $cm_{\mathbb{T}} + (1-c)\nu$ for some $c > 0$ and $\nu \in \mathscr{M}(\mathbb{T})$.
(b) Prove that μ-almost every point $x \in \mathbb{T}$ has a dense orbit under S_2.

Exercise 5.3.3. Show how to choose, for any $n \geqslant 1$ with $\gcd(n, 6) = 1$, a set $S_n \subseteq \{0, \frac{1}{n}, \ldots, \frac{n-1}{n}\}$ that is invariant under both $x \mapsto 2x \pmod 1$ and $x \mapsto 3x \pmod 1$ with $|S_n| \geqslant n^\alpha$ for some fixed $\alpha > 0$. Show that for any $\varepsilon > 0$ there exists some $N(\varepsilon)$ such that for $n \geqslant N(\varepsilon)$ the set S_n is ε-dense in $[0, 1)$.

Notes to Chapter 5

[(32)](Page 133) The treatment here follows Boshernitzan [12] and Furstenberg [46, Sect. IV] closely.
[(33)](Page 136) The argument in Section 5.2 comes from a paper of Einsiedler and Ward [37] and uses half-space entropy ideas from Kitchens and Schmidt [74] and Einsiedler [34]. Related work for $\mathbb{Z}^d$-actions by toral automorphisms has been done by Kalinin and Katok [68], where more refined information is found about joinings and the consequences of the existence of non-trivial joinings.
[(34)](Page 138) Determining which subgroups generate in this sense is closely related to the expansive subdynamical structure of the system, which is introduced by Boyle and Lind [16] and completely determined for commuting automorphisms of compact abelian groups by Einsiedler, Lind, Miles and Ward [35].
[(35)](Page 143) In fact Ledrappier's "three dot" example has many invariant measures: Einsiedler [33] constructs uncountably many closed invariant sets and uncountably many different invariant measures giving positive entropy to any single map in systems of this sort.

[36](Page 143) This is a special case of a more general result due to Schmidt [128], where it is shown that Haar measure is the unique "most mixing" invariant probability measure for a large class of systems of this sort.

[37](Page 144) Other approaches to this result have been found by Feldman [43], Host [61], Lindenstrauss [89], Parry [113] and others. More recently, Bourgain, Lindenstrauss, Michel and Venkatesh [13] have found effective versions of both Rudolph's and Johnson's theorems, relating the size of the assumed positive entropy to effective properties of the invariant measure.

[38](Page 152) In fact a much stronger statement holds: Johnson and Rudolph [67] showed that the weak*-limit is equal to $m_{\mathbb{T}}$.

6

Entropy and Names

In Chapters 1 to 4 we developed the basic properties of measure-theoretic and topological entropy. In this chapter we introduce some more specialized topics in entropy. In Section 6.1 we expand on the discussion in Section 1.4, by examining the behavior of the measure of atoms in $\bigvee_{j=0}^{n-1} T^{-j}\xi$ for a measurable partition ξ, and in particular we prove the important Shannon–McMillan–Breiman theorem[39]. In Section 6.2 we expand on some of the discussion on the connection between entropy and coding from Chapter 1.

6.1 Shannon–McMillan–Breiman Theorem

Let $(X, \mathscr{B}, \mu, T)$ be an ergodic measure-preserving system, and fix a countable partition ξ with $H_\mu(\xi) < \infty$. We have seen that

$$\frac{1}{n} H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right) = \frac{1}{n}\int I_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right) \mathrm{d}\mu \to h_\mu(T,\xi).$$

Thus the information function with respect to ξ on average gives the entropy. In this section we present the Shannon–McMillan–Breiman theorem which states that the information function itself already converges almost everywhere to give the entropy.

It will be useful to think of the pair (T, ξ) as defining *names*. If the partition is $\xi = (A_1, A_2, \ldots)$ then the (ξ, n)-name $\mathbf{w}_n^\xi(x)$ of a point $x \in X$ is the vector $(a_0, a_1, \ldots, a_{n-1})$ with the property that $T^i(x) \in A_{a_i}$ for $0 \leqslant i < n$. The Shannon–McMillan–Breiman theorem says (in the ergodic case) that for almost every x the measure of the set of points sharing the (ξ, n)-name of x decays exponentially at a rate determined by the entropy $h_\mu(T, \xi)$.

Theorem 6.1 (Shannon–McMillan–Breiman). *Let $(X, \mathscr{B}, \mu, T)$ be an invertible ergodic measure-preserving system and let ξ be a countable partition with $H_\mu(\xi) < \infty$. Then*

$$\frac{1}{n} I_\mu \left(\bigvee_{i=0}^{n-1} T^{-i}(\xi) \right)(x) \to h_\mu(T, \xi)$$

almost everywhere and in L^1_μ. *If* $(X, \mathscr{B}, \mu, T)$ *is any invertible measure-preserving system, then*

$$\frac{1}{n} I_\mu \left(\bigvee_{i=0}^{n-1} T^{-i}(\xi) \right)(x) \to h_{\mu^{\mathscr{E}}_x}(T, \xi)$$

almost everywhere and in L^1_μ.

As with the Birkhoff ergodic theorem itself, the proof depends on a maximal inequality that controls deviations in the information function (see equation (6.2)).

The assumption of invertibility is a convenience because it allows us to use Section 4.3. This assumption can be removed using Exercise 4.4.1, but the form given here is sufficient for our needs. Before starting the proof, we recall some material from Chapter 4. Fix an ergodic measure-preserving system $(X, \mathscr{B}, \mu, T)$ and let ξ be a countable partition of X with $H_\mu(\xi) < \infty$. Then by Proposition 4.8, and equation (4.3) in particular,

$$I^* = \sup_{n \geqslant 1} I_\mu \left(\xi \middle| \bigvee_{i=1}^{n} T^{-i}(\xi) \right) \in L^1_\mu. \tag{6.1}$$

Indeed, for any $\lambda \geqslant 0$ and set $A \in \xi$ we have

$$\mu\left(\{ x \in A \mid I^*(x) > \lambda \} \right) \leqslant \mathrm{e}^{-\lambda} \tag{6.2}$$

from the proof of Proposition 4.8, which shows equation (6.1). We now turn to the proof of the Shannon–McMillan–Breiman theorem, under the assumption of ergodicity first.

Proof of Theorem 6.1 assuming ergodicity. Write

$$f_n = I_\mu \left(\xi \middle| \bigvee_{i=1}^{n} T^{-i}(\xi) \right)$$

for $n \geqslant 1$ and $f_0 = I_\mu(\xi)$. Recall from Proposition 1.7 and equation (1.7) that

$$I_\mu(\xi \vee \eta) = I_\mu(\eta \big| \xi) + I_\mu(\xi)$$

and

$$I_\mu(\xi) \circ T = I_\mu(T^{-1}(\xi)).$$

It follows (by induction) that

$$
\begin{aligned}
I_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}(\xi)\right) &= I_\mu\left(\xi\Big|\bigvee_{i=1}^{n-1} T^{-i}(\xi)\right) + I_\mu\left(\bigvee_{i=1}^{n-1} T^{-i}(\xi)\right) \\
&= f_{n-1} + I_\mu\left(\bigvee_{i=0}^{n-2} T^{-i}(\xi)\right)\circ T \\
&= f_{n-1} + f_{n-2}\circ T + \cdots + f_0 \circ T^{n-1} \\
&= \sum_{k=0}^{n-1} f_{n-1-k}\circ T^k. \qquad (6.3)
\end{aligned}
$$

Notice that by Proposition 4.8 the sequence $f_n = I_\mu\left(\xi\big|\bigvee_{i=1}^{n} T^{-i}\xi\right)$ converges almost everywhere and in L^1_μ to $f = I_\mu\left(\xi\big|\bigvee_{i=1}^{\infty} T^{-i}\xi\right)$. Hence we are tempted to replace the average in equation (6.3) by the ergodic average for f, and this will be done with a careful justification. In fact we may write

$$
\begin{aligned}
\frac{1}{n} I_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}(\xi)\right) &= \frac{1}{n}\sum_{k=0}^{n-1} f_{n-1-k}\circ T^k \\
&= \frac{1}{n}\sum_{k=0}^{n-1} f\circ T^k + \frac{1}{n}\sum_{k=0}^{n-1}(f_{n-1-k}-f)\circ T^k. \qquad (6.4)
\end{aligned}
$$

By the Birkhoff ergodic theorem the first average converges almost everywhere and in L^1_μ to

$$
E_\mu\left(f\big|\mathscr{E}\right) = \int I_\mu\left(\xi\Big|\bigvee_{i=1}^{\infty} T^{-i}(\xi)\right)\,\mathrm{d}\mu_x^{\mathscr{E}},
$$

which equals $h_\mu(T,\xi)$ almost everywhere under the assumption that μ is ergodic. We will return to this expression in the non-ergodic case later. Thus we wish to show (without the assumption of ergodicity) that the second average in equation (6.4) converges to 0 almost everywhere and in L^1_μ. Define, for $N \geqslant 1$,

$$
F_N = \sup_{\ell \geqslant N} |f - f_\ell|,
$$

so that, for $n > N$,

$$
\begin{aligned}
&\left|\frac{1}{n}\sum_{k=0}^{n-1}(f_{n-1-k}-f)\circ T^k\right| \qquad (6.5)\\
&\quad\leqslant \frac{1}{n}\sum_{k=0}^{n-N-1} |f_{n-1-k}-f|\circ T^k + \frac{1}{n}\sum_{k=n-N}^{n-1} |f_{n-1-k}-f|\circ T^{-k} \\
&\quad\leqslant \underbrace{\frac{1}{n}\sum_{k=0}^{n-N-1} F_N\circ T^k}_{A(n,N)} + \underbrace{\frac{1}{n}\sum_{k=n-N}^{n-1} |f_{n-1-k}-f|\circ T^k}_{B(n,N)}.
\end{aligned}
$$

By equation (6.1), $|f_{n-1-k} - f| \leqslant I^* + f \in L^1_\mu$. We claim that $B(n,N) \to 0$ as $n \to \infty$ for fixed N. In fact,

$$\begin{aligned} B(n,N) &\leqslant \frac{1}{n}\sum_{k=n-N}^{n-1} (I^* + f)\circ T^k \\ &= \frac{1}{n}\sum_{k=0}^{n-1} (I^* + f)\circ T^k - \left(\frac{n-N}{n}\right)\left(\frac{1}{n-N}\right)\sum_{k=0}^{n-N-1} (I^* + f)\circ T^k \end{aligned}$$

is bounded by the difference of two expressions that converge to the same limit almost everywhere and in L^1_μ by the pointwise ergodic theorem. Hence

$$B(n,N) \to 0$$

as $n \to \infty$ as claimed, for any fixed N. Recall now that $f_n \to f$ as $n \to \infty$, and $f_n, f \leqslant I^* \in L^1_\mu$ so $F_N \leqslant I^*$ and $F_N \to 0$ as $N \to \infty$. By the dominated convergence theorem we also have $F_N \to 0$ in L^1_μ. Fix $\varepsilon > 0$ and choose N so that $\|F_N\|_1 < \varepsilon$, which implies that $\|A(n,N)\|_1 < \varepsilon$ for $n > N$. Choosing n large enough we also get $\|B(n,N)\|_1 < \varepsilon$ by the argument above, and the convergence of the expression in equation (6.5) to 0 in L^1_μ follows.

For pointwise convergence, choose N with $\|F_N\|_1 < \varepsilon^2$ and notice that

$$A(n,N) \leqslant \frac{1}{n}\sum_{k=0}^{n-1} F_N \circ T^k \to E\left(F_N \middle| \mathscr{E}\right)$$

as $n \to \infty$ by the pointwise ergodic theorem. Hence, by equation (6.5),

$$\begin{aligned} &\mu\Big(\Big\{\, x \mid \limsup_{n\to\infty} \Big|\, \frac{1}{n}\sum_{k=0}^{n-1} (f_{n-1-k} - f)\circ T^k \,\Big| > \varepsilon \,\Big\}\Big) \\ &\quad\leqslant \mu\Big(\{x \mid \limsup_{k\to\infty} A(n,N) > \varepsilon/2\}\Big) + \mu\Big(\{x \mid \limsup_{n\to\infty} B(n,N) > \varepsilon/2\}\Big) \\ &\quad\leqslant \mu\Big(\{x \mid E(F_N|\mathscr{E}) > \varepsilon/2\}\Big) \\ &\quad\leqslant \frac{2}{\varepsilon}\int_{\{x|E(F_N|\mathscr{E})>\varepsilon/2\}} E(F_N|\mathscr{E})\,\mathrm{d}\mu \leqslant 2\varepsilon, \end{aligned}$$

showing the pointwise convergence. □

Notice that the hypothesis of ergodicity was only used in order to identify the expression $E\left(I_\mu(\xi|\bigvee_{i=1}^\infty T^{-i}\xi)\middle|\mathscr{E}\right)$ with $h_\mu(T,\xi)$. As we will see, it would be convenient to use monotonicity of the information function with respect to the given σ-algebra, which as we discussed is false in general (see Example 1.8). The next lemma provides a replacement, by showing that the information function satisfies the monotonicity property needed on average.

Lemma 6.2. *Let $(X, \mathscr{B}, \mu, T)$ be a measure-preserving system, let (ξ_n) be a sequence of partitions, and let $\mathscr{A} \subseteq \mathscr{B}$ be any sub-σ-algebra. Then*

$$\limsup_{n\to\infty} \left(\frac{1}{n} I_\mu(\xi_n) - \frac{1}{n} I_\mu(\xi_n \big| \mathscr{A}) \right) \geqslant 0$$

almost everywhere.

Proof. Define

$$f_n = \sum_{A\in\xi_n} \frac{\mu(A)}{\mu_x^{\mathscr{A}}(A)} \chi_A$$

for $n \geqslant 1$. We claim that f_n has expectation $\int f_n \, \mathrm{d}\mu = 1$. In fact

$$\begin{aligned} \int f_n(x) \, \mathrm{d}\mu(x) &= \iint f_n \, \mathrm{d}\mu_x^{\mathscr{A}} \, \mathrm{d}\mu(x) \\ &= \int \sum_{A\in\xi_n} \frac{\mu(A)}{\mu_x^{\mathscr{A}}(A)} \underbrace{\int \chi_A \, \mathrm{d}\mu_x^{\mathscr{A}}}_{\mu_x^{\mathscr{A}}(A)} \, \mathrm{d}\mu(x) = 1. \end{aligned}$$

It follows that for any fixed $\varepsilon > 0$ we have

$$\mu\left(\{x \mid f_n(x) > \mathrm{e}^{\varepsilon n}\}\right) < \mathrm{e}^{-\varepsilon n} \int_{\{x | f_x(x) > \mathrm{e}^{\varepsilon n}\}} f_n \, \mathrm{d}\mu \leqslant \mathrm{e}^{-\varepsilon n},$$

which implies that $f_n(x) < \mathrm{e}^{\varepsilon n}$ almost everywhere, for all but finitely many n, by the Borel–Cantelli lemma. However, this is precisely the claim that for almost every x

$$\frac{1}{n} I_\mu(\xi_n)(x) - \frac{1}{n} I_\mu(\xi_n \big| \mathscr{A})(x) = -\frac{1}{n} \log \frac{\mu(A)}{\mu_x^{\mathscr{A}}(A)} > -\varepsilon$$

for all but finitely many n, as required. □

Proof of Theorem 6.1 without ergodicity. Since

$$H_\mu\left(\xi \big| \mathscr{E}\right) = \int H_{\mu_x^{\mathscr{E}}}(\xi) \, \mathrm{d}\mu \leqslant H_\mu(\xi)$$

by Lemma 4.6 and monotonicity of entropy, we know that ξ is a countable partition with finite entropy with respect to $\mu_x^{\mathscr{E}}$ for almost every ergodic component $\mu_x^{\mathscr{E}}$. Also,

$$I_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi \big| \mathscr{E} \right)(x) = -\log \mu_x^{\mathscr{E}}\left([x]_{\bigvee_{i=0}^{n-1} T^{-i}\xi} \right) = I_{\mu_x^{\mathscr{E}}}\left(\bigvee_{i=0}^{n-1} T^{-i}\xi \right)$$

by definition. Hence the ergodic case of Theorem 6.1 applies to show that

$$\frac{1}{n} I_\mu \left(\bigvee_{i=0}^{n-1} T^{-i}\xi \middle| \mathscr{E} \right)(x) \longrightarrow h_{\mu_x^{\mathscr{E}}}(T,\xi) \tag{6.6}$$

($\mu_x^{\mathscr{E}}$-almost everywhere with respect to almost every ergodic component $\mu_x^{\mathscr{E}}$ and hence) μ-almost everywhere. As $\mathscr{E}$ is part of the Pinsker algebra, we expect that the left-hand side is essentially equal to the information function, which would give the theorem. We will now use Lemma 6.2 to make this more precise.

Recall that we already know that

$$\frac{1}{n} I_\mu \left(\bigvee_{i=0}^{n-1} T^{-i}\xi \right) \longrightarrow E\left(I_\mu \left(\xi \middle| \bigvee_{i=1}^{\infty} T^{-i}\xi \right) \middle| \mathscr{E} \right), \tag{6.7}$$

so by Lemma 6.2 applied with $\mathscr{A} = \mathscr{E}$, we get from the limits in equations (6.6) and (6.7)

$$E\left(I_\mu \left(\xi \middle| \bigvee_{i=1}^{\infty} T^{-i}\xi \right) \middle| \mathscr{E} \right)(x) \geqslant h_{\mu_x^{\mathscr{E}}}(T,\xi) \tag{6.8}$$

almost everywhere. Integrating gives

$$h_\mu(T,\xi) = \int I_\mu \left(\xi \middle| \bigvee_{i=1}^{\infty} T^{-i}\xi \right) \mathrm{d}\mu \geqslant \int h_{\mu_x^{\mathscr{E}}}(T,\xi)\,\mathrm{d}\mu(x), \tag{6.9}$$

but we already know by Theorem 4.23 that equation (6.9) is an equality, so equation (6.8) must be an equality almost everywhere, and the theorem follows. □

In the course of proving the Shannon–McMillan–Breiman theorem we proved the following result, which will be used again.

Lemma 6.3. *Let $(X, \mathscr{B}, \mu, T)$ be an invertible measure-preserving Borel system, and let ξ be a partition with finite entropy. Then*

$$h_{\mu_x^{\mathscr{E}}}(T,\xi) = E\left(I_\mu \left(\xi \middle| \bigvee_{i=1}^{\infty} T^{-i}\xi \right) \middle| \mathscr{E} \right)(x).$$

6.2 Entropy and Coding

Many of the familiar notions of ergodic theory have analogs in information theory: as we have mentioned in Section 1.2, the *elements of a partition* correspond to a set of *symbols*; a *point* in a shift space defined by that partition corresponds to an infinite string of *data*; a *cylinder set* in the shift space corresponds to a *block*; a *shift-invariant measure* corresponds to a *stationary probability*; the *entropy* with respect to the chosen partition corresponds to the

compression ratio; the *Shannon–McMillan–Breiman theorem* (Theorem 6.1) is called the *asymptotic equipartition property*.

While the source coding theorem may be viewed as a theoretical result in ergodic theory, the practical implementation of efficient codes is both of great importance practically and motivates further results in ergodic theory. We end this section by introducing two practical coding techniques. In real applications there are several conflicting demands: optimality, in the sense of minimizing the average length of a code; efficiency, in the sense of minimizing the amount of storage or computation required to implement the code; utility, in the sense of not requiring too much information about the data source.

We begin by analyzing how the code defined on p. 14 measures up to the source coding theorem in a simple example.

Example 6.4. Suppose we have five symbols $1, 2, 3, 4, 5$ with relative frequencies $\frac{1}{3}, \frac{4}{15}, \frac{1}{5}, \frac{2}{15}, \frac{1}{15}$ respectively. Following the algorithm on p. 14, we compute $\ell_1 = \lceil -\log(1/3) \rceil = 2$, $\ell_2 = 3$, $\ell_3 = 2$, $\ell_4 = 3$, and $\ell_5 = 3$. Then, using equation (1.6), we associate symbols to intervals by

$$\begin{aligned} 1 &\mapsto (0, \tfrac{1}{4}) = I(00), \\ 2 &\mapsto (\tfrac{1}{4}, \tfrac{1}{2}) = I(01), \\ 3 &\mapsto (\tfrac{1}{2}, \tfrac{5}{8}) = I(100), \\ 4 &\mapsto (\tfrac{5}{8}, \tfrac{3}{4}) = I(101), \\ 5 &\mapsto (\tfrac{3}{4}, \tfrac{7}{8}) = I(110). \end{aligned}$$

The resulting code s is shown in Table 6.1, and we have $L(s) = 2.80\ldots$.

Table 6.1. A simple code.

symbol	probability	codeword	length
1	$\frac{1}{3}$	00	2
2	$\frac{4}{15}$	01	2
3	$\frac{1}{5}$	100	3
4	$\frac{2}{15}$	101	3
5	$\frac{1}{15}$	110	3

For this example, the lower bound for $L(s)$ from Lemma 1.10 is $2.15\ldots$ and the upper bound from Lemma 1.11 is $3.59\ldots$.

6.2.1 Huffman Coding

This coding method is both practical and efficient in situations where the probability distribution (the relative frequencies of the symbols) is known in

advance. It works by trying to assign short code words to common symbols, and was devised by Huffman [62] in the early 1950s.

We are given k symbols $\{1, 2, \ldots, k\}$ with corresponding probabilities

$$v_1 \leqslant v_2 \leqslant \cdots \leqslant v_k$$

(by rearrangement if necessary), to which we apply the following algorithm.

SYMBOL MERGING: Combine the two symbols of lowest frequency (initially, this will be 1 and 2) to form a new symbol 12 with probability $v_1 + v_2$.

This gives a new set of symbols – initially $\{12, 3, \ldots, k\}$ – and a new list of corresponding probabilities – initially $\{v_1+v_2, v_3, \ldots, v_k\}$. Now rearrange this data and apply SYMBOL MERGING again, and repeat until only two symbols remain. The code is then constructed by drawing a binary tree starting from a single root, with labels 0 for a left branch and 1 for a right branch, describing how to split the final two symbols back into the original symbols.

Example 6.5. Consider again the example in Example 6.4, which we now arrange to have symbols $1, 2, 3, 4, 5$ with probabilities $\frac{1}{15}, \frac{2}{15}, \frac{1}{5}, \frac{4}{15}, \frac{1}{3}$. The sequence obtained by applying SYMBOL MERGING is as follows:

$$\begin{aligned} 1, 2, 3, 4, 5 &\mapsto 12, 3, 4, 5 \quad \text{with probabilities } \tfrac{1}{5}, \tfrac{1}{5}, \tfrac{4}{15}, \tfrac{1}{3}; \\ 12, 3, 4, 5 &\mapsto 123, 4, 5 \quad \text{with probabilities } \tfrac{2}{5}, \tfrac{4}{15}, \tfrac{1}{3}; \\ 4, 5, 123 &\mapsto 45, 123 \quad \text{with probabilities } \tfrac{3}{5}, \tfrac{2}{5}. \end{aligned}$$

The resulting binary tree is shown in Figure 6.1, the resulting code s_H is shown in Table 6.2, and we find $L(s_H) \log 2 = 2.20\ldots$, showing that this code improves on the simple code in Example 6.4.

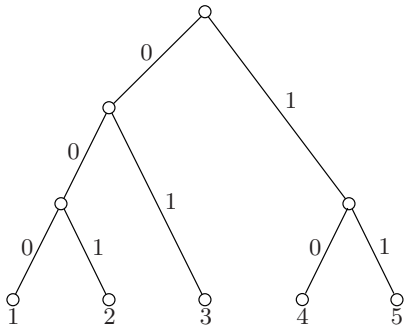

Fig. 6.1. The binary tree for a Huffman code.

Table 6.2. A Huffman code.

symbol	probability	codeword	length
1	$\frac{1}{15}$	000	3
2	$\frac{2}{15}$	001	3
3	$\frac{1}{5}$	01	2
4	$\frac{4}{15}$	10	2
5	$\frac{1}{3}$	11	2

6.2.2 Lempel–Ziv Coding

The Lempel–Ziv algorithm [150] is easy to implement and is widely used because of this. In contrast to Huffman coding, no prior knowledge of the probability distribution is needed. Moreover, Lempel–Ziv coding asymptotically achieves the entropy bound for the compression ratio. For simplicity, we assume that we start with an ergodic measure-preserving system $(X, \mathscr{B}, \mu, T)$ with a partition $\xi = \{A_0, A_1\}$ into two atoms. The resulting process may be identified with the shift space $X = \prod_{i=1}^{\infty}\{0,1\}$ carrying a shift-invariant measure ρ. Thus the source we wish to encode generates binary strings $x_1x_2\ldots$ according to the fixed stationary distribution ρ.

A *parsing* of the string $x_1x_2\ldots$ is a division of the string into blocks

$$w_1|w_2|\ldots.$$

As we have seen in Example 1.9, the definition of a prefix code means that the code of a block can be decoded by parsing the code into codes of symbols.

Lempel–Ziv parsing: Given a binary sequence $x_1x_2\ldots$, parse the sequence inductively as follows. The first block w_1 consists of the single symbol x_1. Now suppose that $w_1|w_2|\cdots|w_j = x_1\ldots x_{n(j)}$, and choose w_{j+1} as follows.

- If $x_{n(j)+1} \notin \{w_1,\ldots,w_j\}$ then set $w_{j+1} = x_{n(j)+1}$.
- If $x_{n(j)+1} \in \{w_1,\ldots,w_j\}$ then set $w_{j+1} = x_{n(j)+1}\ldots x_m$ where $m > n(j)$ is the smallest integer with the property that $x_{n(j)+1}\ldots x_m \in \{w_1,\ldots,w_j\}$ but $x_{n(j)+1}\ldots x_{m+1} \notin \{w_1,\ldots,w_j\}$.

Colloquially, the algorithm defines the next block to be the shortest block that is new (not yet seen). Notice that the parsing scheme is *sequential* in the sense that the choices of blocks do not depend on future symbols. It follows that an initial string of n symbols is parsed as

$$x_1\ldots x_n = w_1|w_2|\cdots|w_m u$$

where the final block u is either empty, or $u = w_j$ for some j, $1 \leqslant j \leqslant m$. Each new block $w = w'x$ is only new because of its final symbol x, so it is determined by recording x (a single binary digit) together with the information about where the block w appeared earlier in the parsed string.

Example 6.6. Starting from the sequence $1011010100010\ldots$, the Lempel–Ziv parsing gives

$$1|0|11|01|010|00|10|\ldots,$$

from which the code constructed reads

$$(000,1)(000,0)(001,1)(010,1)(100,0)(010,0)(001,0)$$

by the following argument:

- $(000,1)$ is the code of 1 because 000 is the location of the empty block;
- $(000,0)$ is the code of 0 similarly;
- $(001,1)$ is the code of 11 because $001 = 1$ is the position of the prefix block 1;
- $(010,1)$ is the code of 01 because $010 = 2$ is the position of the prefix block 0;
- $(100,0)$ is the code of 010 because $100 = 4$ is the position of the prefix block 01, and so on.

The Lempel–Ziv code is asymptotically optimal in the following sense: if c_n denotes the number of blocks in the sequence obtained by parsing an input sequence of length n, then $\log c_n/\log 2$ binary digits are needed to describe the location of the prefix to each block, and one binary digit needed to define the last symbol. Thus the length of the compressed sequence is approximately $c_n(1+\log c_n/\log 2)$, and it may be shown that

$$\frac{1}{n}c_n(1+\log c_n/\log 2)\to h(T,\xi)$$

for almost every sequence. Since we do not pursue the theory further we will not prove this here.

Notes to Chapter 6

[39] (Page 155) This result was proved in increasingly general settings by Shannon [129] (in his development of information theory), McMillan [94], Carleson [21] (extending to countable partitions of finite entropy), Breiman [17], Ionescu Tulcea [63], Chung [24] and Parry [110] (removing the assumption of an invariant measure). As used in ergodic theory it is usually referred to as the Shannon–McMillan–Breiman Theorem.