Chapter 1
Measure-Theoretic Entropy, Introduction

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

attr. von Neumann

Let $(X, \mathscr{B}, \mu)$ be a probability space, and let $T : X \to X$ be a measurable map which we will frequently also refer to as a *transformation*. We say that T is *measure-preserving*, or equivalently that μ is T-*invariant*, if $\mu(T^{-1}B) = \mu(B)$ for every $B \in \mathscr{B}$. In this case we also say that $(X, \mathscr{B}, \mu, T)$ is a *measure-preserving system*. A measure-preserving system is called *ergodic* if a modulo μ invariant set B must have measure $\mu(B) \in \{0, 1\}$, where $B \in \mathscr{B}$ is called *invariant modulo* μ if $\mu(B \triangle T^{-1}B) = 0$. We refer to Appendix A for a brief introduction to these and further concepts of ergodic theory and for some important examples, and refer to [52] for a more thorough background.

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. As we will see in this chapter and even more so in Chapter 3 there is also a close connection to information theory and the pioneering work of Shannon [184] from 1948. The name 'entropy' for Shannon's measure of information carrying capacity was apparently suggested by von Neumann: Shannon is quoted by Tribus and McIrvine [198] 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.' "

The purpose of this volume is to extend this advantage to the reader, who will learn throughout these notes that entropy is a multifaceted notion of

great importance to ergodic theory, dynamical systems, and its applications. For instance entropy can be used in order to distinguish special measures like Haar measure from other invariant measures.

However, let us not jump ahead too much and note that one of the initial motivations for entropy theory was the following kind of question. The *Bernoulli shift* on 2 symbols

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

defined by $(\sigma_{(2)}(x))_n = x_{n+1}$ for every $n \in \mathbb{Z}$ and $x \in \{0,1\}^{\mathbb{Z}}$ preserves the $(\frac{1}{2}, \frac{1}{2})$ Bernoulli measure μ_2 which is defined to be the product measure $\prod_{\mathbb{Z}}(\frac{1}{2}, \frac{1}{2})$ on $\{0,1\}^{\mathbb{Z}}$ (see also Appendix A.4 for more general examples of this type). Similarly, the Bernoulli shift on 3 symbols

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

preserves the $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ Bernoulli measure μ_3. Those two measure-preserving systems share many properties, and in particular are unitarily equivalent in the following sense. To any invertible measure-preserving system $(X, \mathscr{B}, \mu, T)$ one can associate a unitary operator $U_T : L^2_\mu(X)$ defined by $U_T(f) = f \circ T$ for all $f \in L^2_\mu(X)$ (see also Section A.1.1). The two shift maps σ_2 and σ_3 are unitarily equivalent in the sense that 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 $U_{\sigma_2} = WU_{\sigma_3}W^{-1}$ (see Exercise 2.4.4 for a description of a much larger class of measure-preserving systems that are all spectrally indistinguishable).

Are $\sigma_{(2)}$ and $\sigma_{(3)}$ isomorphic as measure-preserving transformations? To see that this is not out of the question, we note that Mešalkin [135] 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 $(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$, see Section 1.7 for a brief description of the isomorphism between these two maps.

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 their interpretation in terms of information theory makes these very easy to remember and highly intuitive.

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, $\xi = \{A_1, \dots, A_k\}$ or $\xi = \{A_1, A_2, \dots\}$. 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, for example, in Sections 1.2, 1.4 and 1.7 it is essential that we work with an enumerated list. 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 η is said to be a *refinement* of ξ, written $\xi \leqslant \eta$ if each atom of ξ is a union of atoms of η. The *common refinement* of two partitions $\xi = \{A_1, A_2, \dots\}$ 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 sub-algebras 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.

For a measurable map $T : X \to X$ and a partition $\xi = \{A_1, A_2, \dots\}$ we 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, 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

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

The formula in (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 $H_\mu(\xi)$ depends on the partition ξ only via the probability vector $(\mu(A_1), \mu(A_2), \dots)$. Restricting to finite probability vectors, the *entropy function* 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 $H_\mu(\cdot\big|\cdot)$. 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 $(\frac{1}{k}, \dots, \frac{1}{k})$.
(4) $H_\mu(\xi \vee \eta) = H_\mu(\eta) + H_\mu(\xi|\eta)$.

Khinchin [102, 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 following monotonicity of slopes

$$\frac{\psi(t) - \psi(s)}{t-s} \leqslant \frac{\psi(u) - \psi(t)}{u-t}. \tag{1.2}$$

Note that strict convexity would give a strict inequality in (1.2).

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*

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

If in addition ψ is strictly convex, then

$$\psi\left(\int f(x)\,\mathrm{d}\mu(x)\right) < \int \psi\left(f(x)\right)\,\mathrm{d}\mu(x) \tag{1.4}$$

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 (1.2),

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

It follows that $\psi(s) \geqslant \psi(t) + (s - t)\beta$ for any s with $a < s < b$, so

$$\psi\left(f(x)\right) - \psi(t) - \left(f(x) - t\right)\beta \geqslant 0 \tag{1.5}$$

for every $x \in X$. Since ψ is continuous,[†] the map $x \mapsto \psi\left(f(x)\right)$ is measurable and so we may integrate (1.5) 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 (1.3).

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 (1.4). □

We note that only (1.2) was needed to prove Lemma 1.3, which shows that $\phi'' \geqslant 0$ implies convexity and $\phi'' > 0$ implies strict convexity only using the mean value theorem of analysis.

We now apply this to the function $x \mapsto x \log x$ in the definition of entropy. Define the function $\phi : [0, \infty) \to \mathbb{R}$ by

$$\phi(x) = \begin{cases} 0 & \text{if } x = 0; \\ x \log x & \text{if } x > 0. \end{cases} \tag{1.6}$$

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 $(0, 1]$, we get the following fundamental lemma.

Lemma 1.4 (Convexity). *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.

[†] On open sub-intervals, this is a consequence of the monotonicity of slopes in (1.2). On half-open intervals, continuity may fail at an end point, but this does not affect measurability.

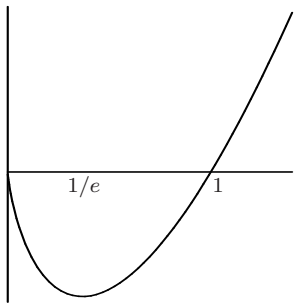

Fig. 1.1: The graph of $x \mapsto \phi(x)$.

Proposition 1.5 (Maximal entropy). *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 page 8. We also note that this proposition is a precursor of many characterizations of 'uniform measures' as being those with 'maximal entropy' (see Example 1.28 for the first instance of this phenomenon).

Proof of Proposition 1.5. By Lemma 1.4, if some atom P has $\mu(P) \neq \frac{1}{k}$, 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 ξ 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

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

In the next proposition we give the remaining main properties of the entropy function, and in particular we prove property (4) from page 8.

Proposition 1.7 (Additivity and Monotonicity). *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) $I_\mu(\xi \vee \eta) = I_\eta(\xi) + I_\mu(\xi|\eta)$, $H_\mu(\xi \vee \eta) = H_\mu(\eta) + H_\mu(\xi|\eta)$ *and so, if $H_\mu(\xi) < \infty$, then*

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

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

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

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 $H_\mu(\xi|\eta) \leqslant H_\mu(\xi)$ for entropy (see property (4) above) does not hold for the information function $I_\mu\left(\cdot\middle|\cdot\right)$. For example, let ξ and η 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

$$I_m(\xi) = \log 2$$

while

$$I_m\left(\xi\middle|\eta\right) \text{ is } \begin{cases} > \log 2 \text{ in the shaded region;} \\ < \log 2 \text{ outside the shaded region.} \end{cases}$$

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

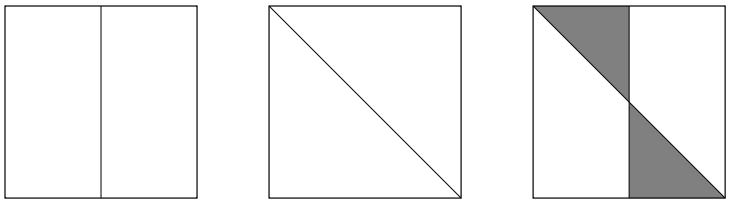

Fig. 1.2: Partitions ξ and η and their refinement.

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

showing the second formula in (1), and hence the first by setting $\eta=\{X\}$.

Notice that

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

which gives (2) by integration.

By convexity of ϕ,

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

showing (3).

Finally, using the above and the already established additivity property (2) we get

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

□

Exercises for Section 1.1

Exercise 1.1.1. Find countably infinite partitions ξ,η of $[0,1]$ with $H_m(\xi)$ finite and with $H_m(\eta)$ infinite, 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 (considered up to sets of measure zero) 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. Extend Proposition 1.7(2) to a conditional form by showing that

$$H_\mu(\xi\vee\eta\big|\,\zeta) = H_\mu(\xi\big|\,\zeta) + H_\mu(\eta\big|\,\xi\vee\zeta) \tag{1.7}$$

for countable partitions ξ,η,ζ.

Exercise 1.1.5. For partitions $\xi = \{A_1,\dots,A_n\}, \eta = \{B_1,\dots,B_n\}$ of fixed cardinality (and thought of as ordered lists), show that $(\xi,\eta)\mapsto H_\mu(\xi\big|\,\eta) = H_\mu(\xi\vee\eta) - H_\mu(\eta)$ is a continuous function of ξ and η with respect to the metric

$$\mathsf{d}(\xi,\eta) = \sum_{i=1}^{n}\mu(A_i\triangle B_i).$$

Exercise 1.1.6. 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 entropy 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 countably infinite partitions.(2) It will be important here to think of a finite or countably infinite 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*, which, (because of the second condition) is also referred to as a *prefix-free code*, is then a map

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

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 $\mathbf{S}$ 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 $\mathbf{S}(i_0)\mathbf{S}(i_1)\dots\mathbf{S}(i_r)$ 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 $\mathbf{S}(1) = 0$, $\mathbf{S}(2) = 10$, $\mathbf{S}(3) = 11$. 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 $10\big|0\big|0\big|11 = \mathbf{S}(2)\mathbf{S}(1)\mathbf{S}(1)\mathbf{S}(3)$.
(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 $|\mathbf{S}(i)|$ denote the length of the codeword $\mathbf{S}(i)$. Then the average length of the code is

$$L(\mathbf{S}) = \sum_i v_i |\mathbf{S}(i)|,$$

which may be finite or infinite depending on the code.

We wish to think of a code $\mathbf{S}$ as a compression algorithm, and in this viewpoint a code $\mathbf{S}$ is better (on average more efficient) than another code $\mathbf{S}'$ if the average length of the code $\mathbf{S}$ is smaller than the average length of the code $\mathbf{S}'$. 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 (Lower bound on average code length). *For any code* $\mathbf{S}$ *the average length satisfies*

$$L(\mathbf{S}) \log 2 \geqslant H(v_1, v_2, \dots) = -\sum_i v_i \log v_i.$$

In other words the entropy $H(v_1, v_2, \dots)$ of a probability vector $(v_1, v_2, \dots)$ gives a lower 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 $\mathbf{S}$ imply Kraft's inequality[(3)]

$$\sum_i 2^{-|\mathbf{S}(i)|} \leqslant 1. \tag{1.8}$$

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

$$I(d_1 \dots d_\ell) = \left(\tfrac{d_1}{2} + \tfrac{d_2}{2^2} + \dots + \tfrac{d_\ell}{2^\ell}, \tfrac{d_1}{2} + \tfrac{d_2}{2^2} + \dots + \tfrac{d_\ell+1}{2^\ell}\right) \tag{1.9}$$

of length $\frac{1}{2^\ell}$. The requirements on the code mean precisely that all the intervals $I(\mathbf{S}(i))$ for $i = 1, 2, \dots$ are disjoint, which proves (1.8). The lemma now follows by convexity of $x \mapsto -\log x$ (see Lemma 1.4):

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

□

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 $\ell_i = \lceil -\log_2 v_i \rceil$ (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 $I(\mathbf{S}(i))$ for a unique address $\mathbf{S}(i) = d_1 \dots d_{|\mathbf{S}(i)|}$ as in (1.9).

Lemma 1.11 (Near optimal code). *Given a probability vector $(v_1, v_2, \dots)$ (permuting the indices if neccessary to assume $v_1 \geqslant v_2 \geqslant \cdots$), there exists a code $\mathbf{S}$, called the* Shannon code*, such that*

$$I_i = \left(\sum_{j=1}^{i-1} \frac{1}{2^{\ell_j}}, \sum_{j=1}^{i} \frac{1}{2^{\ell_j}}\right) = I(\mathbf{S}(i)) = \left(\sum_{k=1}^{|\mathbf{S}(i)|} \frac{d_k}{2^k}, \sum_{k=1}^{|\mathbf{S}(i)|} \frac{d_k}{2^k} + \frac{1}{2^{|\mathbf{S}(i)|}}\right).$$

is the interval with the address $\mathbf{S}(i) = d_1 \cdots d_{|\mathbf{S}(i)|}$. The Shannon code satisfies

$$|\mathbf{S}(i)| = \lceil -\log_2 v_i \rceil$$

and hence $L(\mathbf{S})\log 2 \leqslant H(v_1, v_2, \dots) + \log 2$.

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 (1.9) 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 $I(\mathbf{S}(i))$ for some binary sequence $\mathbf{S}(i)$ of length $|\mathbf{S}(i)|$. The disjointness of the intervals ensures that $\mathbf{S}$ is a code.

The average length of the code $\mathbf{S}$ is, by definition,

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

□

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

Using the Shannon code, we interpret the information (measured by the information function) up to one digit as the number of digits needed to encode a symbol.

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 (Invariance). *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

$$I_\mu\left(\xi \middle| \eta\right) \circ T = I_\mu\left(T^{-1}\xi \middle| T^{-1}\eta\right). \tag{1.10}$$

Proof. It is enough to show (1.10). Notice that $T^{-1}[Tx]_\eta = [x]_{T^{-1}\eta}$ for all x, so

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

□

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 sub-additive sequences is needed.

Lemma 1.13 (Fekete). *Let (a_n) be a sequence of elements of $\mathbb{R} \cup \{-\infty\}$ with the sub-additive property*

$$a_{m+n} \leqslant a_m + a_n$$

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. If $a_n = -\infty$ for some $n \geqslant 1$ then by the sub-additive property we have $a_{n+k} = -\infty$ for all $k \geqslant 1$, so the result holds (with limit $-\infty$).

Assume now that $a_n > -\infty$ for all n, and 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 sub-additive 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 is chosen large enough and we apply division with remainder to write $n = mk + j$, then we may assume 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.

Assume for a moment that 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 the invariance property in Lemma 1.12. In general, subadditivity of entropy (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. This asymptotic linear growth 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*.

By the same argument as above, one sees that the sequence (a_n) defined by

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

is sub-additive in the sense of Lemma 1.13, which shows the claimed convergence and the second equality 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

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

One fact that might explain why entropy is such a useful notion is that there are many possible ways to define entropy. Let us give immediately a second possible definition. As always we ask the reader to find reasonable descriptions of the entropy expressions in terms of information gain (instead of just relying on our formal manipulations of the entropy expressions).

Proposition 1.15 (Entropy conditioned on future). *If $(X, \mathscr{B}, \mu, T)$ is a measure-preserving system and ξ is a countable partition with finite entropy, then*

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

Proof. The limit exists by monotonicity of entropy (Proposition 1.7(4)). By additivity of entropy (Proposition 1.7(2)) we also have for any $n \geqslant 1$ that

$$\begin{aligned} H_\mu\Big(\bigvee_{i=0}^{n-1} T^{-i}\xi\Big) &= H_\mu\left(T^{-(n-1)}\xi\right) + H_\mu\left(T^{-(n-2)}\xi\big|T^{-(n-1)}\xi\right) \\ &\qquad + \cdots + H_\mu\Big(\xi\Big|\bigvee_{i=1}^{n-1} T^{-i}\xi\Big) \\ &= H_\mu(\xi) + H_\mu\left(\xi\big|T^{-1}\xi\right) + \cdots + H_\mu\Big(\xi\Big|\bigvee_{i=1}^{n-1} T^{-i}\xi\Big), \end{aligned}$$

where we used invariance (Lemma 1.12) in the last step. Thus

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

showing the result since the Césaro limit of a convergent sequence coincides with the limit of the original sequence. □

Example 1.16. Let $X_{(2)} = \{0,1\}^{\mathbb{Z}}$ with the Bernoulli $(\frac{1}{2},\frac{1}{2})$ measure $\mu_{(2)}$, preserved by the shift $\sigma_{(2)}$. 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

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

1.3.1 Elementary Properties

Notice that we are not yet in a position to compute $h_{\mu_{(2)}}(\sigma_{(2)})$ from Example 1.16, 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 1.17. *Let $(X, \mathscr{B}, \mu, T)$ be a measure-preserving system on a probability space, and let ξ and η be countable partitions of X with finite entropy. Then we have*

(1) **(Trivial bound)** $h_\mu(T, \xi) \leqslant H_\mu(\xi)$;
(2) **(Subadditivity)** $h_\mu(T, \xi \vee \eta) \leqslant h_\mu(T, \xi) + h_\mu(T, \eta)$;
(3) **(Continuity bound)** $h_\mu(T, \eta) \leqslant h_\mu(T, \xi) + H_\mu(\eta|\xi)$.

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 $n \geqslant 1$,

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

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

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

Subadditivity follows by taking $n \to \infty$.

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

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

by the invariance property in Lemma 1.12. Taking $n \to \infty$ we obtain the continuity bound. □

Proposition 1.18 (Iterates). *Let $(X, \mathscr{B}, \mu, T)$ be a measure-preserving system on a probability space and ξ a countable partition of X with finite entropy. Then*

(1) $h_\mu(T, \xi) = h_\mu\big(T, \bigvee_{i=0}^{k} T^{-i}\xi\big)$ *for all* $k \geqslant 1$;

(2) *for invertible* T*,*

$$h_\mu(T,\xi) = h_\mu(T^{-1},\xi) = h_\mu\Big(T, \bigvee_{i=-k}^{k} T^{-i}\xi\Big)$$

for all $k \geqslant 1$*;*
(3) $h_\mu(T^k) = kh_\mu(T)$ *for* $k \geqslant 1$*; and*
(4) $h_\mu(T) = h_\mu(T^{-1})$ *if* T *is invertible.*

Proof. (1): For any $k \geqslant 1$,

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

(2): For any $n \geqslant 1$, the invariance property (Lemma 1.12) shows that

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

Dividing by n and taking the limit gives the first statement, and the second equality follows easily along the lines of (1).

(3): For any partition ξ with finite entropy,

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

It follows that

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

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

For the reverse inequality, notice that

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

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

(4): This follows from (2). □

Lemma 1.19 (Finite vs. finite entropy). *Entropy can be computed using finite partitions only, in the sense that*

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

In fact, for every countable partition ξ with finite entropy and $\varepsilon > 0$ there exists a finite partition η (measurable with respect to $\sigma(\xi)$) with $H_\mu(\xi|\eta) < \varepsilon$.

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 $H_\mu(\xi) < \infty$. By the continuity bound in Proposition 1.17(3) it suffices to show the last claim in the lemma. To see this, let $\xi = \{A_1, A_2, \dots\}$ and define

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

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 since $\phi(B_N) < 0$, it is possible to choose N large enough to ensure that $H_\mu(\xi\big|\eta) < \varepsilon$. □

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 1.20 (Entropy of factor). *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 1.21 (Kolmogorov–Sinaĭ). *Let $(X, \mathscr{B}, \mu, T)$ be a measure-preserving system on a probability space, and let ξ be a partition of finite entropy that is a* one-sided generator *under T in the sense that*

$$\bigvee_{n=0}^{\infty} T^{-n}\xi = \mathscr{B}. \tag{1.11}$$

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

Theorem 1.21 transfers some of the difficulty inherent in computing entropy onto the problem of finding a generator. We note that if a partition is found satisfying (1.11) modulo μ then (under the assumption that $\mathscr{B}$ is countably generated, which we will have whenever this is used) there is an isomorphic copy of the system for which we have found a generator satisfying (1.11) as stated. There are general results[(5)] 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 1.6 a generator will be found for a non-trivial example, and in Chapter 4 we will give a proof of the existence of finite generators for any finite entropy ergodic system.

Lemma 1.22 (Continuity). *Let $(X, \mathscr{B}, \mu, T)$ be a measure-preserving system, let ξ be a partition satisfying* (1.11)*, and let η be any partition of X with finite entropy. Then*

$$H_\mu\Big(\eta \Big| \bigvee_{i=0}^{n} T^{-i}\xi\Big) \longrightarrow 0$$

as $n \to \infty$.

PROOF. By the last statement in Lemma 1.19 it suffices to consider a finite partition η. By assumption, the partitions $\bigvee_{j=0}^{n} T^{-j}\xi$ for $n = 1, 2, \dots$ together generate $\mathscr{B}$. This in particular shows that for any $\delta > 0$ and $B \in \mathscr{B}$, there exists some $n \geqslant 1$ and some set

$$A \in \sigma\Big(\bigvee_{j=0}^{n} T^{-j}\xi\Big)$$

for which $\mu(A \triangle B) < \delta$. In fact, it is not hard to see that the set of all measurable sets B with this property is a σ-algebra containing $T^{-n}\xi$ for all $n \geqslant 0$, which gives the claim. Alternatively, this follows quickly from the increasing martingale theorem ([52, Th. 5.5]).

Applying the above to all the elements of $\eta = \{B_1, \dots, B_m\}$, we can find one n with the property that there is a collection of sets

$$A_i' \in \sigma\Big(\bigvee_{j=0}^{n} T^{-j}\xi\Big)$$

with $\mu(A_i' \triangle B_i) < \delta/m^2$ for $i = 1, \dots, m-1$. Write

$$A_1 = A_1', A_2 = A_2' \smallsetminus A_1', A_3 = A_3' \smallsetminus (A_1' \cup A_2'), \dots,$$
$$A_{m-1} = A_{m-1}' \smallsetminus \bigcup_{j=1}^{m-2} A_j', \text{ and } A_m = X \smallsetminus \bigcup_{j=1}^{m-1} A_j'.$$

Now notice for $i = 1, \dots, m-1$ that

$$\begin{aligned}\mu(A_i \triangle B_i) &= \mu(A_i \smallsetminus B_i) + \mu(B_i \smallsetminus A_i)\\ &\leqslant \mu(A_i' \smallsetminus B_i) + \mu(B_i \smallsetminus A_i') + \mu\Big(B_i \cap \bigcup_{j=1}^{i-1} A_j'\Big)\\ &\leqslant \frac{\delta}{m^2} + \sum_{j=1}^{i-1} \mu(A_j' \smallsetminus B_j) \leqslant \frac{\delta}{m}\end{aligned}$$

by construction and since $\eta = \{B_1, \dots, B_m\}$ forms a partition. Using that both η and $\zeta = \{A_1, \dots, A_m\}$ form partitions we also get

$$\mu(A_m \triangle B_m) = \mu\Big(\bigcup_{i=1}^{m-1} A_i \triangle \bigcup_{i=1}^{m-1} B_i\Big) \leqslant \sum_{j=1}^{m-1} \mu(A_i \triangle B_i) \leqslant \delta.$$

To summarize, the two partitions η and ζ have the property that

$$\mu(A_i \triangle B_i) < \delta$$

for $i = 1, \dots, m$. Thus

$$H_\mu\Big(\eta\big|\bigvee_{i=0}^{n} T^{-i}\xi\Big) \leqslant H_\mu\left(\eta\middle|\zeta\right) \qquad \text{(by monotonicity (Prop. 1.7(4)))}$$
$$= -\sum_{i=1}^{m} \mu(A_i \cap B_i) \log \frac{\mu(A_i \cap B_i)}{\mu(A_i)}$$
$$- \sum_{i,j=1, i\neq j}^{m} \mu(A_i \cap B_j) \log \frac{\mu(A_i \cap B_j)}{\mu(A_i)}.$$

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

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

as needed. □

Proof of Theorem 1.21. Let ξ be a one-sided generator under T. For any partition η, continuity of entropy (Proposition 1.17 and Lemma 1.22) shows that

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

so $h(T, \eta) \leqslant h(T, \xi)$ as required. The proof for a generator under an invertible T is similar. □

Corollary 1.23. *If $(X, \mathscr{B}, \mu, T)$ is an invertible measure-preserving system on a probability space with a one-sided generator, then $h_\mu(T) = 0$.*

Proof. Let ξ be a partition with

$$\bigvee_{n=0}^{\infty} T^{-n}\xi = \mathscr{B},$$

so that

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

by the Kolmogorov–Sinaĭ theorem (Theorem 1.21) and since entropy can be expressed by conditioning on the future (Proposition 1.15). On the other hand, since T is invertible we may consider the partition $T\xi$ and obtain

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

since T^{-1} preserves the measure and by continuity of entropy (Lemma 1.22). □

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 1.26 and 1.27 the positive entropy reflects the way in which the transformation moves nearby points apart and thereby using the partition chops up the space in a complicated way; in Examples 1.24 and 1.25 the transformation moves points around in a very orderly way, and this is reflected[(6)] in the zero entropy.

Example 1.24. 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 1.25. 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 1.18(3) and Example 1.24. 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 since R_α is just a translation this also gives $x_2 \in R_\alpha^{-n}[0, \frac{1}{2})$ but $x_1 \in R_\alpha^{-n}[\frac{1}{2}, 1)$. This implies that the elements of $\bigvee_{i=0}^{n} T^{-i}\xi$ are intervals whose maximal length decreases to 0 as $n \to \infty$. Therefore the smallest σ-algebra containing $\bigvee_{n=0}^{\infty} T^{-n}\xi$ contains all open sets, and so it follows that $h_{m_\mathbb{T}}(R_\alpha) = 0$ by Corollary 1.23.

Example 1.26. The state partition for the Bernoulli 2-shift in Example 1.16 is a two-sided generator, so we deduce that $h_{\mu_2}(\sigma_2) = \log 2$. In fact $\bigvee_{i=-n}^{n} T^{-i}\xi$ consists of the 2^{2n+1} distinct *cylinder sets*

$$[w]_{-n}^{n} = \{x \in X_{(2)} \mid x_i = w_i \text{ for } i = -n, \ldots, n\}$$

for $w \in X_{(2)}$. It follows that $\bigvee_{n=-\infty}^{\infty} T^{-n}\xi$ contains all metric balls (see Example A.5 for an explicit description of the metric).

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 1.27. 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 $H_{m_{\mathbb{T}}}\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right) = \log 2^n$. The Kolmogorov–Sinaĭ theorem (Theorem 1.21) shows that $h_{m_{\mathbb{T}}}(T_2) = \log 2$.

Example 1.28. Just as in Example 1.27, 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, we have

$$h_\mu(T_p) = h_\mu(T_p, \xi) \leqslant H_\mu(\xi) \leqslant \log p \tag{1.12}$$

by the trivial bound in Proposition 1.17(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 (1.12). 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 again 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 1.28, 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 utilizes the strict convexity of the map $x \mapsto x \log x$ or the map $x \mapsto -\log x$. We will see other instances of this in Chapter 8.

Example 1.29. Let $(X, \mu, \sigma) = (\mathsf{X}_G^{(v)}, \mu_{\mathbf{p},P}, \sigma)$ be the Markov shift defined in Section A.4.2. Then

$$h_\mu(\sigma) = -\sum_{i,j} p_i p_{i,j} \log p_{i,j}.$$

To see this, notice that the state partition $\xi = \{[i]_0\}$ is a generator, so we may apply the Kolmogorov–Sinaĭ theorem (Theorem 1.21). We have

$$\mu\left([i_0]_0 \cap \sigma^{-1}[i_1]_0 \cap \cdots \cap \sigma^{-(n-1)}[i_{n-1}]_0\right) = p_{i_0} p_{i_0,i_1} \cdots p_{i_{n-2},i_{n-1}},$$

which gives the result using the properties of the logarithm, since by assumption we have $\sum_i p_i p_{i,j} = p_j$ for all j and $\sum_j p_{i,j} = 1$ for all i.

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 $h(T)$ can be expressed as $\lim_{n\to\infty} h(T, \xi_n)$.

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 of full support on the shift space $X = \{0,1\}^{\mathbb{Z}}$ with $h_\mu(\sigma) = 0$.

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 $h_\mu(T, \xi)$ by the following steps.

(1) Recall that

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

from the proof of Proposition 1.15, and deduce that

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

(2) Use (1) and additivity of entropy (Proposition 1.7(2)) 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. Consider finite enumerated partitions of $(X, \mathscr{B}, \mu)$ with k elements. Show that $h_\mu(T, \xi)$ is a continuous function of ξ in the L^1_μ norm on ξ.

Exercise 1.3.6. Show that for any $h \in [0, \infty]$, there is an ergodic measure-preserving transformation with entropy h.

Exercise 1.3.7. (7) Let $(X, \mathscr{B}, \mu, T)$ be a measure-preserving system, and let ξ be a finite partition (or a countably infinite partition with $H_\mu(\xi) < \infty$). Prove that

$$h_\mu(T, \xi) = \inf_{F \subseteq \mathbb{N}} \frac{1}{|F|} H_\mu\left(\bigvee_{n \in F} T^{-n} \xi \right),$$

where the infimum is taken over all finite subsets $F \subseteq \mathbb{N}$.

1.4 Defining Entropy using Names

†We mentioned 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.

1.4.1 Decay Rate

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: the logarithmic rate of decay of the volume of the set associated to a typical name is the entropy.

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 3.1. This link between the decay rate and the entropy is the content of the Shannon–McMillan–Breiman theorem (Theorem 3.1).

Lemma 1.30 (Decay for the Bernoulli shift). *Let $(X, \mathscr{B}, \mu, T)$ be the Bernoulli shift defined by the probability vector $\mathbf{p} = (p_1, \ldots, p_s)$, which means that $X = \prod_{\mathbb{Z}} \{1, \ldots, s\}$, $\mu = \prod_{\mathbb{Z}} (p_1, \ldots, p_s)$, and let $T = \sigma$ be the left shift. Let ξ be the state partition defined by the 0th coordinate of the points in X. Then*

† This section has motivational character, both for definitions already made and for upcoming results, but will not be needed later.

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

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

$$\mu\left(\mathbf{w}_n^\xi(x)\right) = p_{x_0} \cdots p_{x_{n-1}}. \tag{1.13}$$

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

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

so we may rearrange (1.13) to obtain

$$\mu\left(\mathbf{w}_n^\xi(x)\right) = p_1^{\sum_{i=0}^{n-1} \mathbb{1}_1(T^i x)} p_2^{\sum_{i=0}^{n-1} \mathbb{1}_2(T^i x)} \cdots p_s^{\sum_{i=0}^{n-1} \mathbb{1}_s(T^i x)}. \tag{1.14}$$

Now, by the ergodic theorem, for any $\varepsilon > 0$ and for almost every $x \in X$ there is an N so that for every $n \geqslant N$ and $j = 1, \ldots, s$ we have

$$\left|\frac{1}{n}\sum_{i=0}^{n-1} \mathbb{1}_j(T^i x) - p_j\right| < \varepsilon. \tag{1.15}$$

Taking the logarithm in (1.14) and dividing by n we see — to within a small error — the familiar entropy formula in Definition 1.1. More precisely, we combine (1.14)–(1.15) and conclude that

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

so

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

as $n \to \infty$. □

1.4.2 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 [180, 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 Theorem 3.1, the definition of topological entropy using open covers in Section 5.2, and Section 6.3).

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 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 3.1) in the form

$$-\frac{1}{n} \log \mu\left(\mathbf{w}_n^\xi(x)\right) \longrightarrow h_{\mu,\mathrm{name}}(T, \xi) \tag{1.16}$$

for μ-almost every x.

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

$$h_{\mu,\mathrm{name}}(T, \xi) = h_{\mu,\mathrm{name}}(T). \tag{1.17}$$

We shall see later that $h_{\mu,\mathrm{name}}(T, \xi) = h_\mu(T, \xi)$ as a corollary of Theorem 3.1 (see Exercise 3.1.2).

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 1.30 (see Exercise 1.4.1).

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

Exercises for Section 1.4

Exercise 1.4.1. Show that $h_{\mu,\mathrm{name}}(\sigma,\xi) = H(\mathbf{p})$ (using both the notation and the statement of Lemma 1.30).

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 Section 1.3. 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 optimal Shannon code s for ξ induces in a natural way a code s_n on names of length n by concatenation,

$$s_n(i_0 \dots i_{n-1}) = s(i_0)s(i_1)\dots s(i_{n-1}). \tag{1.18}$$

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 (1.18), 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 1.31. 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 1.16 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 than $\frac{\log 3}{\log 2}$, 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 $h_{\mu_3}(\sigma_{(3)}) = \log 3$.

We will develop these ideas further in Section 3.2.

Exercises for Section 1.5

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-free code for the blocks of length n over A. Use the source coding theorem in Section 1.2 to show that

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

1.6 An Entropy Calculation for a Group Automorphism

†An illuminating example of a compact group automorphism is the map

$$T = T_A : \mathbb{T}^2 \to \mathbb{T}^2$$

defined by

$$T : \begin{pmatrix} x \\ y \end{pmatrix} \longmapsto \begin{pmatrix} y \\ x+y \end{pmatrix} \pmod 1.$$

This map is associated to the matrix $A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$ in a natural way. Since T is a surjective endomorphism of a compact group, it preserves the Lebesgue measure m on $\mathbb{T}^2$ (see [52, 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. We will show in Section 8.4 in greater generality that the Lebesgue measure can be characterized as the only invariant measure that achieves the maximal value of the entropy for the automorphism.

Theorem 1.32 (Golden mean automorphism). *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 1.32 is a special case of a general result for automorphisms of the torus, which will be shown in Theorem 6.9 by other methods. We will prove[(8)] Theorem 1.32 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 $-1/\rho < 0$.

† While we certainly think it is good to see this example early on, this section is only discussing a particular measure-preserving system and so could be skipped.

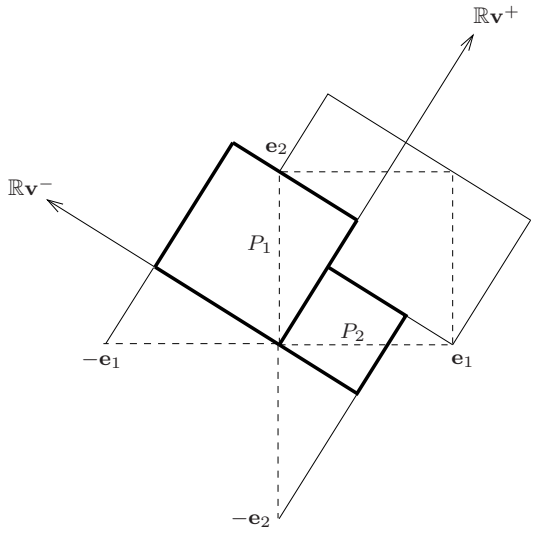

Fig. 1.3: A partition of $\mathbb{T}^2$ adapted to the geometry of the automorphism.

Let $\xi = \{P_1, P_2\}$ denote the partition of $\mathbb{T}^2$ into the two regions shown in Figure 1.3. In Figure 1.3 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 2.7 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 1.4 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.

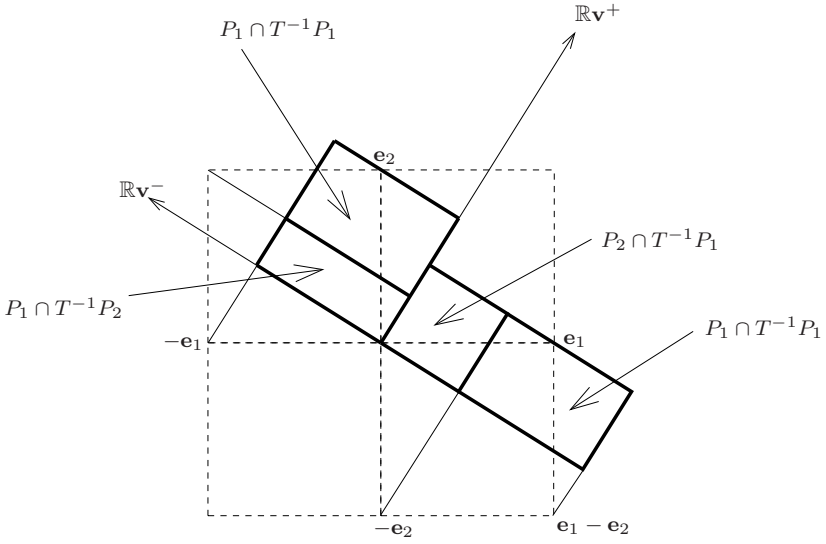

Fig. 1.4: The three rectangles in $\xi \vee T^{-1}\xi$.

Lemma 1.33 (Geometry of partitions). *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 between $\frac{1}{10}\rho^{-n}$ and $2\rho^{-n}$. In particular, ξ is a generator for T.*

Proof. We start by proving the first statement by induction. The discussion before the statement of the lemma and Figure 1.4 comprise the case $n = 1$. We leave it to the reader to check that the lengths of the edges of P_1 and P_2 in the direction of $\mathbf{v}^+$ are indeed between $\frac{1}{10}$ and 2.

Assume therefore that 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 1.5).

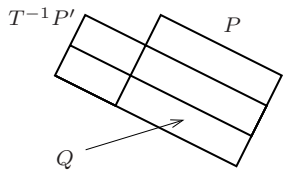

Fig. 1.5: An atom in $\xi \vee T^{-1}\xi \vee \cdots \vee T^{-(n+1)}\xi$.

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

$$P \cap Q = \big(P \cap T^{-1}P'\big) \cap \big(T^{-1}P' \cap Q\big)$$

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.

Recall that ξ is a generator for the invertible map T if

$$\bigvee_{k=-\infty}^{\infty} T^{-k}\xi = \mathscr{B}_{\mathbb{T}^2}. \tag{1.19}$$

To see that this is the case, notice first that the partition elements of

$$\bigvee_{k=-n}^{n} T^{-k}\xi$$

consist of rectangles of diameter at most $c\rho^{-n}$ for some $c > 0$. Therefore, every open set can be written as a union of elements in $\bigvee_{k=-\infty}^{\infty} T^{-k}\xi$, and (1.19) follows. □

By the Kolmogorov–Sinaĭ theorem (Theorem 1.21), Lemma 1.33 reduces the proof of Theorem 1.32 to calculating $h_m(T) = h_m(T, \xi)$.

Proof of Theorem 1.32. By Lemma 1.33, all elements of the partition

$$\xi \vee T^{-1}\xi \vee \cdots \vee T^{-n}\xi$$

have Lebesgue measure in the interval $[c_1\rho^{-n}, c_2\rho^{-n}]$ for some absolute constants $c_1, c_2 > 0$. Using the definition of the information function, this implies that

$$-\log c_2 + n\log\rho \leqslant I_m\left(\xi \vee T^{-1}\xi \vee \cdots \vee T^{-n}\xi\right) \leqslant -\log c_1 + n\log\rho.$$

After dividing by n and letting $n \to \infty$, we see that $h_m(T,\xi) = \log\rho$. By Lemma 1.33 and the Kolmogorov–Sinaĭ theorem (Theorem 1.21), we obtain $h_m(T) = \log\rho$. □

1.7 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,(9) see also the survey article of Weiss [207]. 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 going any further, we mention a simple example of a family of isomorphisms found by Mešalkin.

Example 1.34. As mentioned on page 6, Mešalkin [135] 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 1.26 we can show that

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

Mešalkin showed(10) 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 [89] 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

† The content of this section will not be needed later.

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

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 1.6. View the short blocks 0 as poor people, and the tall blocks as wealthy ones.

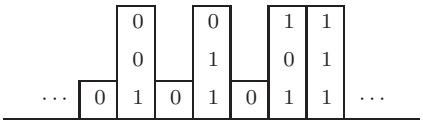

Fig. 1.6: A typical point in the $(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$ Bernoulli shift.

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:

$$\begin{array}{ccc} 0 & & \\ 1 & \longrightarrow & 1\ 0 \\ 1\ 0 & & 1\ 0 \end{array}$$

- 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 neighbourhood, 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 [193]) 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.

Example 1.35. In the same spirit as the construction of Mešalkin above, Kalikow and Weiss [95] found an explicit isomorphism between the full shift on $\prod_{n\in\mathbb{Z}}[0,1]$ with the infinite product of Lebesgue measure, and the full shift on $\prod_{n\in\mathbb{Z}}\mathbb{N}$ with the infinite product of the discrete measure $(p_1, p_2, \ldots)$

for certain probability distributions satisfying the necessary entropy condition

$$-\sum_{n=1}^{\infty} p_i \log p_i = \infty.$$

We refer to their paper [95] for the details of this remarkable construction, which exploits a code similar to that of Example 1.34 in an infinite iterated process.

For the following discussion we need to introduce a slight restriction on the type of measure spaces that we want to consider. A *Borel probability space* is a Borel subset X 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.

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 1.36. 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 any choice of 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 1.37. 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 1.38. *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

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

Measure-preserving systems X and Y are said to be *weakly isomorphic* if each is a factor of the other. Theorem 1.20 really shows that entropy is an invariant of weak isomorphism. It is far from obvious, but true,[(11)] that systems can be weakly isomorphic without being isomorphic. Sinaĭ showed [189] 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 [188] 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 a factor of Y. Thus, for example, Bernoulli shifts of the same entropy are weakly isomorphic. Ornstein's isomorphism theorem (proved in [149] for finite entropy and extended to the infinite entropy case in [150]) 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 1.38 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 1.39. (1) The automorphism of $\mathbb{T}^2$ associated to the matrix $\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$ (see Section 1.6) is a Bernoulli automorphism.

(2) More generally, Katznelson [100] 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 [122] 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 [118] and Miles and Thomas [136], [138], [137]. Some simplifications were made by Aoki [9].
(4) A mixing Markov shift is a Bernoulli automorphism (see Ornstein and Shields [152].)
(5) Certain ergodic automorphisms of nilmanifolds are Bernoulli automorphisms (see Dani [39]).
(6) The map of geodesic flow for a fixed time on a surface of negative curvature is a Bernoulli automorphism (see Ornstein and Weiss [154]).
(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 [67]).
(8) A generalization of (5) is that any mixing Anosov flow preserving a smooth measure is a Bernoulli automorphism (see Ratner [170] or Bunimovič [30]).
(9) The β-transformation $T_\beta : [0,1] \to [0,1]$ is defined for each $\beta > 1$ by $T_\beta(x) = \beta x$ modulo 1. There is a T_β-invariant measure μ_β on $[0,1]$ absolutely continuous with respect to Lebesgue measure, discovered by Rényi [172]. Then the invertible extension of the system (T_β, μ_β) is a Bernoulli automorphism (see [52, Ex. 2.1.7] for the details of the invertible extension construction, and Smorodinsky [192] or Fischer [59] for the result).
(10) Notice that a Bernoulli automorphism automatically has positive entropy (we exclude the map on a single point). It follows that a zero entropy system (for example, a rotation on a compact group, the horocycle flow for a fixed time, or a unipotent flow for a fixed time on a homogeneous space) is never isomorphic to a Bernoulli automorphism.

The definitive nature of Ornstein's Theorem should not mask the scale of the problem of classifying measure-preserving transformations up to isomorphism in general: Bernoulli shifts are a significant class, encompassing many geometrically natural maps, but the structure of most measure-preserving systems remains mysterious.[(12)]

Notes to Chapter 1

[(1)] (Page 5) The original material may be found in papers of Kolmogorov [108] (corrected in [107]), Rokhlin [175], and Rokhlin and Sinaĭ [178]. For an attractive survey of the foundations and later history of entropy in ergodic theory, see the survey article by Katok [98].

The concept of entropy is due originally to the physicist Clausius [37], who used it in connection with the dispersal of usable energy in thermodynamics in 1854 and coined the term 'entropy' in 1868. Boltzmann [18] 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 [184] in information theory.

(2) (Page 15) The connections between information theory and ergodic theory, many of which originate with the work of Shannon [184], are pervasive (these will be discussed further in Sections 3.1 and 3.2).

(3) (Page 16) This inequality, and the converse result that if a list of integers $\ell_1, \ell_2 \dots$ satisfies the inequality (1.8) then there is a prefix-free code with $\ell_i = |\mathbf{S}(i)|$, was obtained by Kraft [109] and McMillan [133].

(4) (Page 19) This seems to have first been proved by Fekete [56, p. 233] (in multiplicative form); a more accessible source is Pólya and Szegő [167, Chap. 3, Sect. 1].

(5) (Page 25) The main result concerning the existence of generators is due to Krieger [90]: 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 1.17(1) it is not possible for there to be a generator with fewer atoms, so this result is optimal.

(6) (Page 28) 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.

(7) (Page 31) This holds more generally for measure-preserving actions of amenable groups, as stated in Ollagnier [141, Sec. 4.3].

(8) (Page 36) These geometrically natural generators were introduced in work of Adler and Weiss [6], [7].

(9) (Page 40). The theory described in this section is due to Ornstein, and it is outlined in his monograph [151]. An elegant treatment using joinings may be found in the monograph of Rudolph [180].

(10) (Page 40) 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 1.20, the additional necessary condition that the two shifts have the same entropy).

(11) (Page 43) This question was answered in the thesis of Polit [166], who constructed a pair of weakly isomorphic transformations of zero entropy that are not isomorphic. Rudolph [179] 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 [197] using Gaussian processes, and by Lemańczyk [117] using product cocycles. More recently, Kwiatkowski, Lemańczyk, and Rudolph [113] have constructed weakly isomorphic C^∞ volume-preserving diffeomorphisms of $\mathbb{T}^2$ that are not isomorphic.

(12) (Page 44) 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 [52, 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 [57] interpreted a construction of many mutually non-isomorphic K-automorphisms by Ornstein and Shields [153] 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 [62] have used Hjorth's theory of turbulent equivalence relations [84] 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.