Chapter 3
Entropy and Names

In Chapters 1–2 we developed the basic properties of measure-theoretic entropy. In Section 3.1 we expand on the discussion in Section 1.4, by examining the behavior of the measure of atoms in ξ_0^{n-1} for a measurable partition ξ, and in particular we prove the important Shannon–McMillan–Breiman theorem,[(18)] which in a sense is the ergodic theorem of entropy theory. In Section 3.2 and 3.3 we expand on some of the discussion on the connection between entropy and coding from Chapter 1. In Section 3.4 we present an analog of the Shannon–McMillan–Breiman theorem for Bowen balls, in which we use a metric to define dynamical neighbourhoods instead of partitions.

3.1 Shannon–McMillan–Breiman Theorem

Let $(X, \mathscr{B}, \mu, T)$ be an ergodic measure-preserving system, and fix a countable partition ξ with $H_\mu(\xi) < \infty$. We have seen that

$$\frac{1}{n} H_\mu\left(\xi_0^{n-1}\right) = \frac{1}{n} \int I_\mu\left(\xi_0^{n-1}\right) \, \mathrm{d}\mu \to h_\mu(T, \xi).$$

Thus the information function with respect to ξ on average gives the entropy. In this section we present the Shannon–McMillan–Breiman theorem which states that the information function itself already converges almost everywhere to give the entropy.

It will be useful to think of the pair (T, ξ) as defining *names*. If the partition is $\xi = (A_1, A_2, \ldots)$ then the (ξ, n)-name $\mathbf{w}_n^\xi(x)$ of a point $x \in X$ is the vector $(a_0, a_1, \ldots, a_{n-1})$ with the property that $T^i(x) \in A_{a_i}$ for $0 \leqslant i < n$. The Shannon–McMillan–Breiman theorem says (in the ergodic case) that for almost every x the measure of the set of points sharing the (ξ, n)-name of x decays exponentially at a rate determined by the entropy $h_\mu(T, \xi)$.

Theorem 3.1 (Shannon–McMillan–Breiman). *Let $(X, \mathscr{B}, \mu, T)$ be an invertible ergodic measure-preserving system and let ξ be a countable partition with $H_\mu(\xi) < \infty$. Then*

$$\frac{1}{n} I_\mu\left(\xi_0^{n-1}\right)(x) \to h_\mu(T, \xi)$$

almost everywhere and in L^1_μ.

As with the Birkhoff ergodic theorem itself, the proof depends on a maximal inequality that controls deviations in the information function (see (3.1)). The assumption of invertibility is a convenience because it allows us to use Section 2.4. This assumption can be removed using Exercise 2.6.1, but the form given here is sufficient for our needs.

We will also prove the following more general version, where we drop the assumption of ergodicity and consider relative entropy over an invariant factor.

Theorem 3.2 (Relative Shannon–McMillan–Breiman). *Let $\mathscr{A} \subseteq \mathscr{B}$ be an invariant σ-algebra in an invertible system $(X, \mathscr{B}, \mu, T)$, and let ξ be a countable partition with $H_\mu(\xi) < \infty$. Then*

$$\frac{1}{n} I_\mu\left(\xi_0^{n-1} \middle| \mathscr{A}\right)(x) \to h_{\mu_x^{\mathscr{E}}}(T, \xi \middle| \mathscr{A})$$

almost everywhere and in L^1_μ.

Before starting the proofs, we recall some material from Chapter 2. Fix a measure-preserving system $(X, \mathscr{B}, \mu, T)$ and let ξ be a countable partition of X with $H_\mu(\xi) < \infty$. Also let $\mathscr{A}$ be a T-invariant σ-algebra. Then by Proposition 2.14, and (2.3) in particular,

$$I^* = \sup_{n \geqslant 1} I_\mu\left(\xi \middle| \xi_1^n \vee \mathscr{A}\right) \in L^1_\mu. \tag{3.1}$$

We now turn to the proof of the Shannon–McMillan–Breiman theorem, under the assumption of ergodicity first. We will include the σ-algebra $\mathscr{A}$ even in the proof of Theorem 3.1 since we will use this also as the initial part of the proof of Theorem 3.2.

Proof of Theorem 3.1 (relative to $\mathscr{A}$). Write

$$f_n = I_\mu\left(\xi \middle| \xi_1^n \vee \mathscr{A}\right)$$

for $n \geqslant 1$ and $f_0 = I_\mu(\xi|\mathscr{A})$. Recall the additivity formula (Proposition 2.13(1))

$$I_\mu(\xi \vee \eta \big| \mathscr{A}) = I_\mu(\xi \big| \eta \vee \mathscr{A}) + I_\mu(\eta \big| \mathscr{A})$$

for any countable partition η with finite entropy, and the invariance property (Lemma 2.17) which we may state as

$$I_\mu(\xi\big|\mathscr{A})\circ T = I_\mu(T^{-1}\xi\big|\mathscr{A})$$

by invariance of $\mathscr{A}$. It follows (by induction) that

$$\begin{aligned} I_\mu\left(\xi_0^{n-1}\big|\mathscr{A}\right) &= I_\mu\left(\xi\big|\xi_1^{n-1}\vee\mathscr{A}\right) + I_\mu\left(\xi_1^{n-1}\big|\mathscr{A}\right) \\ &= f_{n-1} + I_\mu\left(\xi_0^{n-2}\big|\mathscr{A}\right)\circ T \\ &= f_{n-1} + f_{n-2}\circ T + \cdots + f_0\circ T^{n-1} \\ &= \sum_{k=0}^{n-1} f_{n-1-k}\circ T^k \end{aligned} \tag{3.2}$$

for any $n \geqslant 1$. Notice that by Proposition 2.14 the sequence (f_n) defined by $f_n = I_\mu\left(\xi\big|\xi_1^n\vee\mathscr{A}\right)$ for all $n\geqslant 1$ converges almost everywhere and in L^1_μ to $f = I_\mu\left(\xi\big|\xi_1^\infty\vee\mathscr{A}\right)$. Hence we are tempted to replace the average in (3.2) by the ergodic average for f, and this will be done with a careful justification. In fact we may write

$$\begin{aligned} \frac{1}{n} I_\mu\left(\xi_0^{n-1}\big|\mathscr{A}\right) &= \frac{1}{n}\sum_{k=0}^{n-1} f_{n-1-k}\circ T^k \\ &= \frac{1}{n}\sum_{k=0}^{n-1} f\circ T^k + \frac{1}{n}\sum_{k=0}^{n-1}(f_{n-1-k}-f)\circ T^k \end{aligned} \tag{3.3}$$

for all $n\geqslant 1$. By the Birkhoff ergodic theorem the first average converges almost everywhere and in L^1_μ to

$$E_\mu\left(f\big|\mathscr{E}\right)(x) = \int I_\mu\left(\xi\big|\xi_1^\infty\vee\mathscr{A}\right)\,\mathrm{d}\mu_x^{\mathscr{E}}, \tag{3.4}$$

where in the ergodic case $\mu_x^{\mathscr{E}} = \mu$ and by the future formula for entropy (Proposition 2.19(1)) we obtain

$$E_\mu\left(f\big|\mathscr{E}\right) = H_\mu\left(\xi\big|\xi_1^\infty\vee\mathscr{A}\right) = h_\mu(T,\xi\big|\mathscr{A})$$

almost everywhere. We will return to the expression in (3.4) in the non-ergodic case later.

Thus we wish to show (ergodicity will not be needed for that step) that the second average in (3.3) converges to 0 almost everywhere and in L^1_μ. Define, for $N\geqslant 1$,

$$F_N = \sup_{\ell\geqslant N}|f - f_\ell|,$$

so that, for $n > N$,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{k=0}^{n-1}(f_{n-1-k}-f)\circ T^k\right| \qquad (3.5)\\
&\quad\leqslant \frac{1}{n}\sum_{k=0}^{n-N-1}|f_{n-1-k}-f|\circ T^k+\frac{1}{n}\sum_{k=n-N}^{n-1}|f_{n-1-k}-f|\circ T^{-k}\\
&\quad\leqslant \underbrace{\frac{1}{n}\sum_{k=0}^{n-N-1}F_N\circ T^k}_{A(n,N)}+\underbrace{\frac{1}{n}\sum_{k=n-N}^{n-1}|f_{n-1-k}-f|\circ T^k}_{B(n,N)}.
\end{aligned}$$

By (3.1), $|f_{n-1-k}-f|\leqslant I^*+f\in L^1_\mu$. We claim that $B(n,N)\to 0$ as $n\to\infty$ for fixed N. In fact,

$$\begin{aligned}
B(n,N)&\leqslant \frac{1}{n}\sum_{k=n-N}^{n-1}(I^*+f)\circ T^k\\
&=\frac{1}{n}\sum_{k=0}^{n-1}(I^*+f)\circ T^k-\left(\frac{n-N}{n}\right)\left(\frac{1}{n-N}\right)\sum_{k=0}^{n-N-1}(I^*+f)\circ T^k
\end{aligned}$$

is bounded by the difference of two expressions that converge to the same limit almost everywhere and in L^1_μ by the pointwise ergodic theorem. Hence

$$B(n,N)\to 0$$

as $n\to\infty$ as claimed almost everywhere and in L^1_μ, for any fixed N.

Recall now that $f_n\to f$ as $n\to\infty$, and $f_n, f\leqslant I^*\in L^1_\mu$ so

$$F_N\leqslant I^*$$

and $F_N\to 0$ as $N\to\infty$. By the dominated convergence theorem we also have $F_N\to 0$ in L^1_μ. Fix $\varepsilon>0$ and choose N so that $\|F_N\|_1<\varepsilon$, which implies that $\|A(n,N)\|_1<\varepsilon$ for $n>N$. Choosing n large enough we also get $\|B(n,N)\|_1<\varepsilon$ by the argument above, and the convergence of the expression in (3.5) to 0 in L^1_μ follows.

For pointwise convergence, choose N with $\|F_N\|_1<\varepsilon^2$ and notice that

$$A(n,N)\leqslant\frac{1}{n}\sum_{k=0}^{n-1}F_N\circ T^k\to E_\mu\left(F_N\middle|\mathscr{E}\right)$$

as $n\to\infty$ almost everywhere by the pointwise ergodic theorem. Hence, by (3.5) and the above argument for $B(n,N)$ we obtain

$$\begin{aligned}
&\mu\Big(\Big\{\, x \mid \limsup_{n\to\infty} \Big|\, \frac{1}{n}\sum_{k=0}^{n-1} (f_{n-1-k} - f)\circ T^k \,\Big| > \varepsilon \,\Big\}\Big) \\
&\quad\leqslant \mu\Big(\{x \mid \limsup_{k\to\infty} A(n,N) > \varepsilon/2\}\Big) + \mu\Big(\{x \mid \limsup_{n\to\infty} B(n,N) > \varepsilon/2\}\Big) \\
&\quad\leqslant \mu\Big(\{x \mid E_\mu(F_N \big| \mathscr{E}) > \varepsilon/2\}\Big) \\
&\quad\leqslant \frac{2}{\varepsilon}\int_{\{x\mid E_\mu(F_N\mid\mathscr{E})>\varepsilon/2\}} E_\mu(F_N\big|\mathscr{E})\,\mathrm{d}\mu \leqslant 2\varepsilon,
\end{aligned}$$

showing the pointwise convergence. □

We note again that the hypothesis of ergodicity was only used in order to identify the expression $E_\mu\left(I_\mu(\xi\big|\xi_1^\infty \vee \mathscr{A})\big|\mathscr{E}\right)$ with $h_\mu(T,\xi\big|\mathscr{A})$. As we will see, it would be convenient to use monotonicity of the information function with respect to the given σ-algebra, which as we discussed is false in general (see Example 1.8). The next lemma provides a replacement, by showing that the information function satisfies the monotonicity property needed on average.

Lemma 3.3 (Monotonicity of information on average). *Let $(X,\mathscr{B},\mu)$ be a Borel probability space, let (ξ_n) be a sequence of partitions, and let $\mathscr{A},\mathscr{C}$ be sub-σ-algebras of $\mathscr{B}$. Then*

$$\liminf_{n\to\infty} \frac{1}{n}\left(I_\mu(\xi_n\big|\mathscr{A}) - I_\mu(\xi_n\big|\mathscr{C}\vee\mathscr{A})\right) \geqslant 0$$

almost everywhere.

Proof. Define

$$f_n = \sum_{P\in\xi_n} \frac{\mu_x^{\mathscr{A}}(P)}{\mu_x^{\mathscr{C}\vee\mathscr{A}}(P)}\mathbb{1}_P$$

for $n \geqslant 1$. We claim that f_n has expectation $\int f_n\,\mathrm{d}\mu = 1$. In fact

$$\begin{aligned}
\int f_n(x)\,\mathrm{d}\mu(x) &= \iint f_n\,\mathrm{d}\mu_x^{\mathscr{C}\vee\mathscr{A}}\,\mathrm{d}\mu(x) \\
&= \int \sum_{P\in\xi_n} \frac{\mu_x^{\mathscr{A}}(P)}{\mu_x^{\mathscr{C}\vee\mathscr{A}}(P)} \underbrace{\int \mathbb{1}_P\,\mathrm{d}\mu_x^{\mathscr{C}\vee\mathscr{A}}}_{\mu_x^{\mathscr{C}\vee\mathscr{A}}(P)}\,\mathrm{d}\mu(x) = 1.
\end{aligned}$$

It follows that for any fixed $\varepsilon > 0$ we have

$$\mu\big(\{x \mid f_n(x) > \mathrm{e}^{\varepsilon n}\}\big) < \mathrm{e}^{-\varepsilon n}\int_{\{x\mid f_x(x)>\mathrm{e}^{\varepsilon n}\}} f_n\,\mathrm{d}\mu \leqslant \mathrm{e}^{-\varepsilon n},$$

which implies that $f_n(x) < \mathrm{e}^{\varepsilon n}$ almost everywhere, for all but finitely many n, by the Borel–Cantelli lemma. However, this is precisely the claim that for

almost every x

$$\frac{1}{n} I_\mu(\xi_n \big| \mathscr{A})(x) - \frac{1}{n} I_\mu(\xi_n \big| \mathscr{C} \vee \mathscr{A})(x) = -\frac{1}{n} \log f_n(x) > -\varepsilon$$

for all but finitely many n, as required. □

Proof of Theorem 3.2. Recall that the proof of Theorem 3.1 (see (3.3) and (3.4)) gives

$$\frac{1}{n} I_\mu \left(\xi_0^{n-1} \big| \mathscr{A} \right) \longrightarrow E_\mu \left(I_\mu \left(\xi \big| \xi_1^\infty \vee \mathscr{A} \right) \big| \mathscr{E} \right), \tag{3.6}$$

We also claim that

$$\frac{1}{n} I_\mu \left(\xi_0^{n-1} \big| \mathscr{E} \vee \mathscr{A} \right)(x) \longrightarrow h_{\mu_x^{\mathscr{E}}}(T, \xi \big| \mathscr{A}). \tag{3.7}$$

Let us assume the claim for now and use the above monotonicity on average to conclude the proof of the theorem. Indeed, combining Lemma 3.3 with (3.6) and (3.7) we obtain

$$h_{\mu_x^{\mathscr{E}}}(T, \xi \big| \mathscr{A}) \leqslant E_\mu \left(I_\mu \left(\xi \big| \xi_1^\infty \vee \mathscr{A} \right) \big| \mathscr{E} \right)(x) \tag{3.8}$$

almost everywhere. Taking the integral over x gives

$$h_\mu(T, \xi \big| \mathscr{A}) = \int h_{\mu_x^{\mathscr{E}}}(T, \xi \big| \mathscr{A}) \, \mathrm{d}\mu(x) \leqslant H_\mu \left(\xi \big| \xi_1^\infty \vee \mathscr{A} \right) = h_\mu(T, \xi \big| \mathscr{A})$$

by the entropy formula regarding the ergodic decomposition (Theorem 2.33) and the future formula for entropy (Proposition 2.19(1)). This shows that (3.8) must be an equality almost surely. Combining this equality with (3.6)–(3.7) the theorem follows.

Hence it remains to prove the claim in (3.7). For this we will apply the conclusion of Theorem 3.1 to the ergodic components of μ. Since

$$H_\mu \left(\xi \big| \mathscr{E} \right) = \int H_{\mu_z^{\mathscr{E}}}(\xi) \, \mathrm{d}\mu \leqslant H_\mu(\xi)$$

by Lemma 2.11 and monotonicity of entropy, we know that ξ is a countable partition with finite entropy with respect to $\mu_z^{\mathscr{E}}$ for almost every ergodic component $\mu_z^{\mathscr{E}}$. Also,

$$\begin{aligned} I_\mu \left(\xi_0^{n-1} \big| \mathscr{E} \vee \mathscr{A} \right)(x) &= -\log \mu_x^{\mathscr{E} \vee \mathscr{A}} \left([x]_{\xi_0^{n-1}} \right) \\ &= -\log (\mu_z^{\mathscr{E}})_x^{\mathscr{A}} \left([x]_{\xi_0^{n-1}} \right) \\ &= I_{\mu_z^{\mathscr{E}}} \left(\xi_0^{n-1} \big| \mathscr{A} \right)(x) \end{aligned}$$

for almost every z and $\mu_z^{\mathscr{E}}$-almost every x by definition and the double conditioning† formula (Proposition 2.4). Hence Theorem 3.1 applied to the ergodic measure $\mu_z^{\mathscr{E}}$ (and conditioned on $\mathscr{A}$ as proven above) gives (3.7) ($\mu_z^{\mathscr{E}}$-almost everywhere with respect to almost every ergodic component $\mu_z^{\mathscr{E}}$ and hence) μ-almost everywhere. This shows the claim and the theorem. □

In the course of proving the Shannon–McMillan–Breiman theorem we proved the following result, which will be used again.

Lemma 3.4 (Entropy of ergodic component). *Let $(X, \mathscr{B}, \mu, T)$ be an invertible measure-preserving Borel system, and let ξ be a partition with finite entropy. Then $h_{\mu_x^{\mathscr{E}}}(T, \xi) = E_\mu\big(I_\mu\big(\xi\big|\xi_1^\infty\big)\big|\mathscr{E}\big)(x)$.*

Exercises for Section 3.1

Exercise 3.1.1 (Abramov's formula [2]). Let $(X, \mathscr{B}, \mu, T)$ be an invertible ergodic measure-preserving system on a Borel probability space, and let A be any $\mathscr{B}$-measurable set with positive measure. Define (as in Exercise A.1.2) the first return map (also known as the *induced* map) by

$$T_A(x) = T^{r_A(x)}(x)$$

for almost every $x \in A$, where

$$r_A(x) = \min\{n \geqslant 1 \mid T^m(x) \in A\}$$

is the first return time.
(a) Show that the partition ρ_A of A into *level sets* of r_A (that is, into the sets of the form $A_k = \{x \in A \mid r_A(x) = k\}$) has finite entropy with respect to the restricted normalized measure $\mu_A = \frac{1}{\mu(A)}\mu|_A$.
(b) Use Theorem 3.1 to prove the following formula of Abramov. Let ξ be a partition of A with finite entropy with respect to μ_A which refines ρ_A. Then

$$h_{\mu_A}(T_A, \xi) = \frac{1}{\mu(A)} h_\mu(T, \widetilde{\xi}),$$

where $\widetilde{\xi} = \xi \cup \{X \setminus A\}$.
(c) Deduce that $h_{\mu_A}(T_A) = \frac{1}{\mu(A)} h_\mu(T)$.

Exercise 3.1.2. Let $(X, \mathscr{B}, \mu, T)$ be an invertible ergodic measure-preserving system on a Borel probability space. Let ξ be a countable partition with finite entropy. Show the following claims (which establish the link to the notion of name entropy mentioned in Section 1.4).
(a) Let $\varepsilon \in (0, 1)$. Show that

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

where $N(n, \varepsilon)$ is the minimal number of partition elements in ξ_0^{n-1} that are needed to cover a set of μ-measure greater than ε.

† Note that with respect to $\mu_z^{\mathscr{E}}$ the σ-algebra $\mathscr{E}$ is trivial and so $\mathscr{A} \underset{\mu_z^{\mathscr{E}}}{=} \mathscr{A} \vee \mathscr{E}$.

(b) Drop the assumption of ergodicity and describe the meaning of

$$\lim_{\varepsilon\to 1}\limsup_{n\to\infty}\frac{1}{n}\log N(n,\varepsilon)=\lim_{\varepsilon\to 1}\liminf_{n\to\infty}\frac{1}{n}\log N(n,\varepsilon).$$

3.2 Entropy and Coding

Many of the familiar notions of ergodic theory have analogs in information theory.[19] As we have mentioned in Section 1.2, the *elements of a partition* correspond to a set of *symbols*; a *point* in a shift space defined by that partition corresponds to an infinite string of *data*; a *cylinder set* in the shift space corresponds to a *block*; a *shift-invariant measure* corresponds to a *stationary probability*; the *Shannon–McMillan–Breiman theorem* (Theorem 3.1) is often used in a weaker form (namely for a Bernoulli measure and with convergence in probability) and is then called the *asymptotic equipartition property*. The most important notion, that we have already seen as a barrier on the efficiency of coding, is entropy. The *entropy* of a process (that is, of the behavior of a measurable partition under iteration of a measure-preserving transformation) is a theoretical lower bound on the average *compression ratio* (seen in Lemma 1.10, and in a stronger form in Section 3.3). Recall also that a near-optimal compression algorithm exist (by Lemma 1.11), and that this pair of assertions together comprise the *source coding theorem*.

While the source coding theorem may be viewed as a theoretical result in ergodic theory, the practical implementation of efficient codes is both of great importance practically and motivates further results in ergodic theory. We introduce now two practical coding techniques. In real applications there are several conflicting demands: optimality, in the sense of minimizing the average length of a code; efficiency, in the sense of minimizing the amount of storage or computation required to implement the code; utility, in the sense of not requiring too much information about the data source.

We begin by analyzing how the code defined on page 17 measures up to the source coding theorem in a simple example.

Example 3.5. Suppose we have five symbols $1,2,3,4,5$ with relative frequencies $\frac{1}{3}, \frac{4}{15}, \frac{1}{5}, \frac{2}{15}, \frac{1}{15}$ respectively. Following the algorithm on page 17, we compute $\ell_1 = \lceil -\log(1/3)\rceil = 2$, $\ell_2 = 2$, $\ell_3 = 3$, $\ell_4 = 3$, and $\ell_5 = 3$. Then, using (1.9), we associate symbols to intervals by

$$\begin{aligned}
1 &\mapsto (0,\tfrac{1}{4}) = I(00),\\
2 &\mapsto (\tfrac{1}{4},\tfrac{1}{2}) = I(01),\\
3 &\mapsto (\tfrac{1}{2},\tfrac{5}{8}) = I(100),\\
4 &\mapsto (\tfrac{5}{8},\tfrac{3}{4}) = I(101),\\
5 &\mapsto (\tfrac{3}{4},\tfrac{7}{8}) = I(110).
\end{aligned}$$

The resulting code s is shown in Table 3.1, and we have $L(s) = \frac{36}{15} = 2.4$.

Table 3.1: A simple code.

symbol	probability	codeword	length
1	$\frac{1}{3}$	00	2
2	$\frac{4}{15}$	01	2
3	$\frac{1}{5}$	100	3
4	$\frac{2}{15}$	101	3
5	$\frac{1}{15}$	110	3

For this example, the lower bound for $L(s)$ from Lemma 1.10 is $2.149\ldots$ and the upper bound from Lemma 1.11 is $3.149\ldots$.

3.2.1 Huffman Coding

The Huffman coding method is both practical and efficient in situations where the probability distribution (the relative frequencies of the symbols) is known in advance. It works by trying to assign short code words to common symbols, and was devised by Huffman [86] in the early 1950s.

We are given k symbols $\{1, 2, \ldots, k\}$ with corresponding probabilities

$$v_1 \leqslant v_2 \leqslant \cdots \leqslant v_k$$

(by rearrangement if necessary), to which we apply the following algorithm.

SYMBOL MERGING: Combine the two symbols of lowest frequency (initially, this will be 1 and 2) to form a new symbol 12 with probability $v_1 + v_2$.

This gives a new set of symbols — initially $\{12, 3, \ldots, k\}$ — and a new list of corresponding probabilities — initially $\{v_1 + v_2, v_3, \ldots, v_k\}$. Now rearrange this data and apply SYMBOL MERGING again, and repeat until only two symbols remain. The code is then constructed by drawing a binary tree starting from a single root, with labels 0 for a left branch and 1 for a right branch, describing how to split the final two symbols back into the original symbols.

Example 3.6. Consider again the example in Example 3.5, which we now arrange to have symbols $1, 2, 3, 4, 5$ with probabilities $\frac{1}{15}, \frac{2}{15}, \frac{1}{5}, \frac{4}{15}, \frac{1}{3}$. The sequence obtained by applying SYMBOL MERGING is as follows:

$$\begin{aligned}
1, 2, 3, 4, 5 &\mapsto 12, 3, 4, 5 \quad \text{with probabilities } \tfrac{1}{5}, \tfrac{1}{5}, \tfrac{4}{15}, \tfrac{1}{3};\\
12, 3, 4, 5 &\mapsto 123, 4, 5 \quad \text{with probabilities } \tfrac{2}{5}, \tfrac{4}{15}, \tfrac{1}{3};\\
4, 5, 123 &\mapsto 45, 123 \quad \text{with probabilities } \tfrac{3}{5}, \tfrac{2}{5}.
\end{aligned}$$

The resulting binary tree is shown in Figure 3.1, the resulting code s_H is shown in Table 3.2, and we find $L(s_H) = \frac{35}{15} = 2.2$, showing that this code improves on the simple code in Example 3.5.

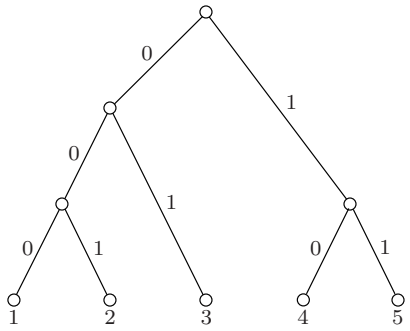

Fig. 3.1: The binary tree for a Huffman code.

Table 3.2: A Huffman code.

symbol	probability	codeword	length
1	$\frac{1}{15}$	000	3
2	$\frac{2}{15}$	001	3
3	$\frac{1}{5}$	01	2
4	$\frac{4}{15}$	10	2
5	$\frac{1}{3}$	11	2

3.2.2 Lempel–Ziv Coding I

The Lempel–Ziv algorithm [215] is easy to implement and is widely used because of this. In contrast to Huffman coding, no prior knowledge of the probability distribution is needed. Moreover, Lempel–Ziv coding asymptotically achieves the entropy bound for the compression ratio. For simplicity, we assume that we start with an ergodic measure-preserving system $(X, \mathscr{B}, \mu, T)$ with a partition $\xi = \{A_0, A_1\}$ into two atoms. The resulting process may be identified with the shift space $X = \prod_{i=1}^{\infty}\{0,1\}$ carrying a shift-invariant measure ρ. Thus the source we wish to encode generates binary strings $x_1x_2\ldots$ according to the fixed stationary distribution ρ.

A *parsing* of the string $x_1x_2\ldots$ is a division of the string into blocks

$$w_1|w_2|\cdots$$

where $w_1, w_2, \ldots$ are finite blocks of symbols $0, 1$ which will also be called *words*. As we have seen in Example 1.9, the definition of a prefix-free code means that the code of a block can be decoded by parsing the code into codes of symbols.

Lempel–Ziv parsing: Given a binary sequence $x_1x_2\ldots$, parse the sequence (that is, break the sequence into consecutive blocks as above) inductively as follows. The first block w_1 consists of the single symbol x_1. Now suppose that $w_1|w_2|\cdots|w_j = x_1\ldots x_{n(j)}$, and choose w_{j+1} as follows.

- If $x_{n(j)+1} \notin \{w_1, \ldots, w_j\}$ then set $w_{j+1} = x_{n(j)+1}$.
- If $x_{n(j)+1} \in \{w_1, \ldots, w_j\}$ then set

$$w_{j+1} = x_{n(j)+1}\ldots x_{m+1},$$

where $m > n(j)$ is the smallest integer with the property that

$$x_{n(j)+1}\ldots x_m \in \{w_1, \ldots, w_j\}$$

but

$$x_{n(j)+1}\ldots x_{m+1} \notin \{w_1, \ldots, w_j\}.$$

Colloquially, the algorithm defines the next block to be the shortest block that is new (not yet seen). Notice that the parsing scheme is *sequential* in the sense that the choices of blocks do not depend on future symbols.

Example 3.7. Starting from the sequence $1011010100010\ldots$, the Lempel–Ziv parsing gives

$$1|0|11|01|010|00|10|\ldots.$$

Notice that in the Lempel–Ziv parsing, each new block $w = w'x$ is only new because of its final symbol x, so it is determined by recording x (a single binary digit) together with the information about where the block w' appeared earlier in the parsed string.

3.2.3 Lempel–Ziv Coding II

Given the parsed sequence

$$w_1|w_2|w_3|\cdots$$

with $w_1 = x_1$ and $w_{j+1} = x_{n(j)+1}\cdots x_{n(j+1)} = w_{\ell(j+1)}x_{n(j+1)}$ for all $j \geqslant 1$, we recode this sequence into the list of pairs

$$(0, x_1)|(\ell(2), x_2)|(\ell(3), x_3)|\cdots$$

where each block w_k is replaced by the address $\ell(k)$ of the earlier block with which w_k starts, followed by its final symbol (that makes it a new block). Here the address 0 is used for the empty block.

Example 3.8. Returning to Example 3.7, the parsed sequence

$$1|0|11|01|010|00|10|\ldots,$$

is recoded to read

$$(0,1)|(0,0)|(1,1)|(2,1)|(4,0)|(2,0)|(1,0)|\cdots$$

In order to end up with a binary code an additional step is needed. For this, notice that by construction $0 \leqslant \ell(k) < k$ so that we need $\lceil \frac{\log k}{\log 2} \rceil$ binary digits to express $\ell(k)$.

3.2.4 Lempel–Ziv Coding III

Given the recoded list of pairs

$$(0,x_1)|(\ell(2),x_2)|(\ell(3),x_3)|\cdots \tag{3.9}$$

we can express each of the integers $\ell(k)$ as a binary block $[\ell(k)]_m$ of length precisely $m = \lceil \frac{\log k}{\log 2} \rceil$ representing $\ell(k)$ in binary. In this way the sequence (3.9) is recoded into the sequence

$$0x_1[\ell(2)]_1 x_{n(2)}[\ell(3)]_2 x_{n(3)}\cdots$$

which we will refer to as the *Lempel–Ziv code* of $x_1x_2x_3\cdots$.

Example 3.9. Returning to Example 3.7 once again, the sequence

$$(0,1)|(0,0)|(1,1)|(2,1)|(4,0)|(2,0)|(1,0)|\cdots$$

is recoded to the binary sequence

$$(0,1)|(0,0)|(01,1)|(10,1)|(100,0)|\cdots$$

where the parentheses and the parsing separators | are included for convenience only. The final Lempel–Ziv coding is

$$01000111011000\cdots.$$

Notice that the original sequence can be sequentially reconstructed from the Lempel–Ziv code. In fact, the length of digits used for each individual block w_n is completely determined by n. Thus the separators and the parentheses can be reintroduced into the code, giving back the list of pairs in (3.9).

By induction, this list then determines the words w_n and thence the original sequence $x_1x_2x_3\cdots$.

The Lempel–Ziv code is asymptotically optimal, but we will not pursue this here (see [215] for the details).

3.3 Shannon–McMillan–Breiman Theorem in Coding Form

In this section we will extend the source coding theorem (Lemmas 1.10 and 1.11) which describes the non-dynamical entropy of a partition up to a small error as the average coding length. In fact Theorem 3.10 shows that the dynamical entropy of a partition precisely describes the optimal average coding length per unit of time. For this we will use the prefix-free codes defined in Section 1.2 and discussed further in Section 3.2, where the symbols of the input correspond to a partition of X. Instead of introducing another set of letters for the alphabet, we will just write $\mathbf{C} : \xi \to \bigcup_{\ell\geqslant 1}\{0,1\}^\ell$ for such a code.

Theorem 3.10 (Entropy and effectivity of codes). *Let $(X, \mathscr{B}, \mu, T)$ be an invertible measure-preserving system, and let ξ be a countable partition of X with $H_\mu(\xi) < \infty$. Then the function*

$$h(x) = h_{\mu_x^{\mathscr{B}}}(T, \xi)$$

is characterized by the following two properties.

(a) *For any sequence of prefix-free codes $(\mathbf{C}_n)$ with $\mathbf{C}_n$ defined on ξ_0^{n-1} for $n \geqslant 1$ we have*

$$\liminf_{n\to\infty} \frac{|\mathbf{C}_n([x]_{\xi_0^{n-1}})|}{n} \geqslant \frac{1}{\log 2} h(x) \tag{3.10}$$

almost everywhere.

(b) *There exists a sequence of prefix-free codes $(\mathbf{C}_n)$ with $\mathbf{C}_n$ defined on ξ_0^{n-1} for $n \geqslant 1$ for which*

$$\lim_{n\to\infty} \frac{|\mathbf{C}_n([x]_{\xi_0^{n-1}})|}{n} = \frac{1}{\log 2} h(x) \tag{3.11}$$

almost everywhere.

For brevity we will usually write $\mathbf{C}_n(x)$ for $\mathbf{C}_n([x]_{\xi_0^{n-1}})$ when the context makes it clear what is meant. Also notice that the codes $\mathbf{C}_n$ constructed in the proof of Theorem 3.10(b) are allowed to depend on X, μ, ξ and T but not on the point $x \in X$.

As we will see, the first part of the above theorem will follow from the easy half of the Borel–Cantelli lemma combined with the Shannon–McMillan–Breiman theorem.

Proof of Theorem 3.10(a). Let $(\mathbf{C}_n)$ be any sequence of prefix-free codes, with $\mathbf{C}_n$ defined on ξ_0^{n-1}. We wish to show (a). To do this, fix $\delta > 0$ and $A > 0$, and define the 'level sets'

$$M_j = \{x \in X \mid j\delta \leqslant h(x) < (j+1)\delta\}$$

of the function $h(x) = h_{\mu_x^{\mathscr{E}}}(T, \xi)$ with $j = 1, \dots, \lceil A/\delta \rceil$. Also define the set

$$Y_n = \{x \in X \mid (\log 2)|\mathbf{C}_n(x)| \leqslant (h(x) - 3\delta)n\},$$

where the coding length is too short by at least $3\delta n$, and the set

$$Z_n = \left\{x \in X \mid \mu\big([x]_{\xi_0^{n-1}}\big) \leqslant \mathrm{e}^{-(h(x)-\delta)n}\right\},$$

on which the refined partition atoms are (up to a possible error of size δ) as small as predicted by the entropy. We note that $Y_n \cap M_j$ is empty for $j = 0, 1, 2$ and all $n \geqslant 1$ by definition of Y_n and M_j (and since the code length is always positive). As $\mathbf{C}_n$ is prefix-free there can only be 2^a partition elements of ξ_0^{n-1} which have a $\mathbf{C}_n$-code of length not exceeding a. We set $a = (j-2)\delta n / \log 2$ and use this to see that the number of partition elements in ξ_0^{n-1} needed to cover $Y_n \cap M_j$ is at most $\mathrm{e}^{(j-2)\delta n}$ for $j = 3, \dots, \lceil A/\delta \rceil$. Therefore we obtain

$$\begin{aligned}\mu\big(Y_n \cap Z_n \cap \{x \mid h(x) < A\}\big) &\leqslant \sum_{j=3}^{\lceil A/\delta \rceil} \mu\left(Y_n \cap Z_n \cap M_j\right) \\ &\leqslant \sum_{j=3}^{\lceil A/\delta \rceil} \mathrm{e}^{(j-2)\delta n}\mathrm{e}^{-(j-1)\delta n} = \lceil A/\delta \rceil \mathrm{e}^{-\delta n}.\end{aligned}$$

Thus, by the Borel–Cantelli lemma, almost every $x \in X$ with $h(x) < A$ has

$$x \notin Y_n \cap Z_n$$

for all but finitely many $n \geqslant 1$. However, by the Shannon–McMillan–Breiman theorem (Theorem 3.2) we know that almost every $x \in X$ has $x \in Z_n$ for all but finitely many $n \geqslant 1$. It follows, since A is an arbitrary constant, that for almost every $x \in X$, we have $x \notin Y_n$ for all but finitely many $n \geqslant 1$. Since $\delta > 0$ was also arbitrary, we deduce (3.10) as required. □

For the second part, we use the Shannon code from Lemma 1.11 together with the Shannon–McMillan–Breiman theorem.

Proof of Theorem 3.10(b). For every $n \geqslant 1$ we let $\mathbf{C}_n$ be the Shannon code from Lemma 1.11 for the partition ξ_0^{n-1}, so that

$$-\log_2 \mu\big([x]_{\xi_0^{n-1}}\big) \leqslant |\mathbf{C}_n\big([x]_{\xi_0^{n-1}}\big)| \leqslant -\log_2 \mu\big([x]_{\xi_0^{n-1}}\big) + 1.$$

Dividing by n and taking the limit gives

$$\lim_{n\to\infty} \frac{1}{n}|\mathbf{C}_n\left([x]_{\xi_0^{n-1}}\right)| = \lim_{n\to\infty} \frac{1}{n\log 2} I_\mu(\xi_0^{n-1})(x) = \frac{1}{\log 2} h_{\mu_x^{\mathscr{E}}}(T,\xi)$$

as claimed. □

Theorem 3.10 is indeed a strengthening of the Shannon–McMillan–Breiman theorem (Theorem 3.1), in the sense that Theorem 3.1 can be quickly deduced from it by arguing along the following lines.

Assume that

$$\liminf_{n\to\infty} \frac{-\log \mu\left([x]_{\xi_0^{n-1}}\right)}{n} < h_{\mu_x^{\mathscr{E}}}(T,\xi)$$

on a set of positive measure. Then the Shannon code $\mathbf{S}_n$ for ξ_0^{n-1} from Lemma 1.11, which has the property that

$$|\mathbf{S}_n([x]_{\xi_0^{n-1}})| \leqslant -\log_2 \mu([x]_{\xi_0^{n-1}}) + 1,$$

gives a contradiction to Theorem 3.10(a). The proof of the reverse inequality is similar to the proof of Theorem 3.10(a) above (and uses Theorem 3.10(b)).

Exercises for Section 3.3

Exercise 3.3.1. Give a detailed proof that Theorem 3.10 implies Theorem 3.2.

3.4 Local Entropy

†In this section we will present the analog of the Shannon–McMillan–Breiman theorem for Bowen balls. This local entropy theorem was proved by Brin and Katok [29].

Definition 3.11. Let $T : X \to X$ be a homeomorphism of a compact metric space (X, d). Then the n-Bowen ball[(20)] of radius $r > 0$ at $x \in X$ (also called a Bowen (n, r) ball) is the set

$$D_T(x, n, r) = \big\{y \in X \mid \mathsf{d}(T^k x, T^k y) < r \text{ for } k = 0, \ldots, n-1\big\}.$$

† It may be helpful to skip this for now, and first read Chapter 5 as it is more familiar, and in particular does not need the material here but nonetheless should help to motivate some of the arguments.

Theorem 3.12 (Local entropy — decay rates of Bowen balls). *Let (X, d) be a compact metric space and let $T : X \to X$ be a homeomorphism. For each $x \in X$, $r > 0$ and $n \geqslant 1$, define functions $\underline{h}$ and $\overline{h}$ by*

$$\underline{h}(x, r) = \liminf_{n\to\infty} \frac{-\log \mu(D_T(x, n, r))}{n}$$

and

$$\overline{h}(x, r) = \limsup_{n\to\infty} \frac{-\log \mu(D_T(x, n, r))}{n}.$$

Then

$$\lim_{r\to 0} \underline{h}(x, r) = \lim_{r\to 0} \overline{h}(x, r) = h_{\mu_x^{\mathscr{E}}}(T)$$

almost everywhere with respect to μ.

Proof. Fix $r > 0$ and suppose that ξ is a finite partition with

$$\max_{P\in\xi} \operatorname{diam}(P) < r.$$

Then for all $x \in X$ we have

$$[x]_{\xi_0^{n-1}} \subseteq D_T(x, n, r)$$

by definition, so $D_T(x, n, r)$ cannot be smaller in μ-measure than the atom of the partition ξ_0^{n-1} containing x. By the Shannon–McMillan–Breiman theorem, this shows that

$$\overline{h}(x, r) \leqslant h_{\mu_x^{\mathscr{E}}}(T, \xi) \leqslant h_{\mu_x^{\mathscr{E}}}(T)$$

for μ-almost every x, which shows that

$$\lim_{r\to 0} \overline{h}(x, r) \leqslant h_{\mu_x^{\mathscr{E}}}(T)$$

(the limit in r exists simply because the expression is monotone in r as $r \to 0$).

For the second half of the theorem we need to show that the μ-measure of the Bowen ball cannot be much bigger than the size predicted by the entropy function $h_{\mu_x^{\mathscr{E}}}(T)$. To do this we will use the coding reformulation of the Shannon–McMillan–Breiman theorem in Theorem 3.10. Roughly speaking, if the μ-measure of the Bowen balls were much too big, then it is impossible for there to be many disjoint Bowen balls. From this assumption, one can select a maximal disjoint collection of Bowen balls and use them to obtain a coding that is asymptotically better than the known constraint given by entropy. The existence of this super-optimal coding is impossible, giving the contradiction to the assumption.

We isolate a small technical step which will be used again later.

Lemma 3.13 (Small partition with a null boundary). *Let X be a compact metric space, μ a probability measure on X, and $\delta > 0$. Then there exists a finite partition $\xi = \{P_1, \dots, P_k\}$ of X such that* $\operatorname{diam}(P_j) < \delta$ *and $\mu(\partial P_j) = 0$ for $j = 1, \dots, k$.*

Proof. To construct such a partition, choose for each $x \in X$ an $\varepsilon_x \in (0, \delta/2)$ with $\mu\left(\partial B_{\varepsilon_x}(x)\right) = 0$ and note that ε_x can be chosen as a continuity point of the monotone function $\varepsilon \in (0, \delta/2) \mapsto \mu(B_\varepsilon(x))$. In this way we obtain an open cover

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

which has a finite subcover by compactness. From this the partition ξ may be readily constructed. □

Fix $\varepsilon > 0$ and let ξ be a finite partition such that

$$h_{\mu_x^{\mathscr{E}}}(T, \xi) > h_{\mu_x^{\mathscr{E}}}(T) - \varepsilon \tag{3.12}$$

if $h_{\mu_x^{\mathscr{E}}}(T) < \infty$, and

$$h_{\mu_x^{\mathscr{E}}}(T, \xi) > \frac{1}{\varepsilon} \tag{3.13}$$

if $h_{\mu_x^{\mathscr{E}}}(T) = \infty$, holds on a set Y_0 with $\mu(Y_0) > 1 - \varepsilon$. Moreover, we can also demand that

$$\mu(\partial P) = 0 \tag{3.14}$$

for all $P \in \xi$ by Lemma 3.13. Such a partition certainly exists, as one can start with a sequence (ξ_ℓ) with $\xi_\ell \subseteq \sigma(\xi_{\ell+1})$ of finite partitions satisfying (3.14) with

$$\max_{P \in \xi} \operatorname{diam}(P) < \frac{1}{\ell}.$$

Then by the Kolmogorov–Sinaĭ theorem (Theorem 2.20), we see that

$$h_{\mu_x^{\mathscr{E}}}(T, \xi_\ell) \longrightarrow h_{\mu_x^{\mathscr{E}}}(T)$$

for almost every ergodic component.

We will need to control the measure of the set on which a small metric error might cause a decoding error (that is, an ambiguous assignment to partition elements). We therefore define[†]

$$\partial_r(\xi) = \left\{x \in X \mid B_r(x) \not\subseteq P \text{ for any } P \in \xi\right\}.$$

Given ξ and Y_0 as above we now choose r_0 such that for all $r \in (0, r_0)$ the set

$$Y = \left\{x \in Y_0 \mid \mu_x^{\mathscr{E}}\left(\partial_{4r}\xi\right) \lceil \log_2 |\xi| \rceil < \varepsilon\right\} \tag{3.15}$$

[†] It is tempting to view this as a neighbourhood of the boundary of the partition elements, but in a general metric setting the boundaries might be empty while $\partial_r(\xi)$ might be non-empty for some $r > 0$.

has $\mu(Y) > 1-2\varepsilon$. This is certainly possible since $\partial_r(\xi) \subseteq \partial_s(\xi)$ for $s \geqslant r > 0$, and

$$\bigcap_{r>0} \partial_r(\xi) = \bigcup_{P\in\xi} \partial P$$

is a null set by (3.14).

Fix $r \in (0, r_0)$, $A > 0$ and $\delta > 0$. Define the level sets

$$M_j = \big\{x \in X \mid j\delta \leqslant \underline{h}(x,r) < (j+1)\delta \text{ and } \underline{h}(x,r) < A\big\},$$

so that

$$\bigcup_{j=0}^{\lfloor A/\delta \rfloor} M_j = \{x \in X \mid \underline{h}(x,r) < A\}.$$

Now let $n \geqslant 1$ and define $Z(j,n)$ to be a maximal collection of points $z \in X$ with the properties:

(1) $z \in M_j$;
(2) $\mu\left(D_T(z,n,r)\right) > \mathrm{e}^{-(j+2)\delta n}$; and
(3) for $z_1 \neq z_2$ in $Z(j,n)$, the Bowen balls $D_T(z_1,n,r)$ and $D_T(z_2,n,r)$ are disjoint.

By (2) and (3) we have

$$a(j,n) = |Z(j,n)| < \mathrm{e}^{(j+2)\delta n}.$$

Choose an ordering $Z(j,n) = \big\{z_1^{(j,n)}, \ldots, z_{a(j,n)}^{(j,n)}\big\}$. By maximality of the collection of centers $Z(j,n)$ we have

$$\bigcup_{\ell=1}^{a(j,n)} D_T(z_\ell^{(j,n)}, n, 2r) \supseteq \{x \in M_j \mid \mu\left(D_T(z,n,r)\right) > \mathrm{e}^{-(j+2)\delta n}\}, \tag{3.16}$$

for if this inclusion does not hold then we could simply add the missing point to $Z(j,n)$, which was assumed to be maximal. Now let $\mathbf{C}_n^{\text{given}}$ be a given prefix-free codes on ξ_0^{n-1}. We are going to construct a new code $\mathbf{C}_n^{A,\delta}$ using the collection of sets constructed above, and then apply Theorem 3.10(a) to the sequence of codes $\big(\mathbf{C}_n^{A,\delta}\big)$.

Fix $x \in X$. We define $\widetilde{\mathbf{C}}_n^{A,\delta}(x)$ to be

$$0\mathbf{C}_n^{\text{given}}(x) \tag{3.17}$$

if $\underline{h}(x,r) \geqslant A$ or if $x \in M_j$ for some $j \in \{0, \ldots, \lceil A/\delta \rceil\}$ but[†]

$$x \notin \bigcup_{\ell=1}^{a(j,n)} D_T(z_\ell^{(j,n)}, n, 2r).$$

[†] Since $\underline{h}$ is defined as a limit inferior this possibility may happen for infinitely many $n \geqslant 1$.

Suppose now that $\underline{h}(x,r) < A$,

$$x \in M_j \cap D_T(z_\ell^{(j,n)}, n, 2r)$$

with $j \in \{0, \dots, \lceil A/\delta \rceil\}$ determined by the definition of the level sets M_j and $\ell \in \{1, \dots, a(j,n)\}$ chosen minimally with that property. Using these choices, we define the code $\widetilde{\mathbf{C}}_n^{A,\delta}(x)$ to be

$$1[j]_{m_1}[\ell]_{m_2(j)}\mathbf{s}(x), \tag{3.18}$$

where as before $[\cdot]_m$ denotes the binary expansion of a given integer between 0 and $2^m - 1$ of length m, $m_1 = \lceil \log_2(A/\delta) \rceil$, and

$$m_2(j) = \left\lceil \frac{(j+2)\delta n}{\log 2} \right\rceil.$$

Before defining $\mathbf{s}(x)$ we make a few comments regarding the information transmitted by the code in the initial string containing j and ℓ. If we have

$$x \in D_T(z_\ell^{(j,n)}, n, 2r)$$

and $0 \leqslant k < n$, then one of the following must hold:

- $T^k z_\ell \in \partial_{2r}\xi$ (in which case $T^k x \in \partial_{4r}\xi$), or
- $T^k\left(D_T(z_\ell, n, 2r)\right) \subseteq P$ for some $P \in \xi$.

In other words, unless $T^k z_\ell \in \partial_{2r}\xi$, the number ℓ (which is transmitted in the initial string) determines the atom of $T^k x$ with respect to ξ. The remaining string $\mathbf{s}(x)$ will be used to transmit the missing information regarding the atom of $T^k x$ with respect to ξ for all those k for which $T^k z_\ell \in \partial_{2r}\xi$. Let $\mathbf{C}$ be a code for ξ with $|\mathbf{C}(P)| = \lceil \log_2 |\xi| \rceil$ for any $P \in \xi$, and define $\mathbf{s}(x)$ to be the concatenation of $\mathbf{C}(T^k x)$ for those k with $T^k z_\ell \in \partial_{2r}\xi$.

With this additional string $\mathbf{s}(x)$, the receiver then knows the atom $[x]_{\xi_0^{n-1}}$ (and also when the string ends). That is, $\widetilde{\mathbf{C}}_n^{A,\delta}$ is a prefix-free code for some partition ζ_n that refines ξ_0^{n-1}. We note that this code $\widetilde{\mathbf{C}}_n^{A,\delta}$ for ζ_n can be used to define a code $\mathbf{C}_n^{A,\delta}$ for ξ_0^{n-1} with equal or shorter code length by simply choosing from every code appearing on an atom of ξ_0^{n-1} one of the shortest ones.

We now need to analyze the lower growth rate of the coding length on $\mathbf{C}_n^{A,\delta}$. By Theorem 3.10(a) we have

$$h_{\mu_x^{\mathscr{E}}}(T,\xi) \leqslant \liminf_{n\to\infty} \frac{|\mathbf{C}_n^{A,\delta}|}{n} \leqslant \liminf_{n\to\infty} \left(\frac{(j+2)\delta}{\log 2} + \frac{|\mathbf{s}(x)|}{n} \right)$$

for almost every x with $\underline{h}(x,r) < A$, where j is chosen so that $x \in M_j$. In fact the latter inequality holds since, by definition of $\underline{h}$, we must have

$$\frac{-\log \mu(D_T(x,n,r))}{n} < \underline{h}(x,r) + \delta,$$

or equivalently

$$\mu(D_T(x,n,r)) > \mathrm{e}^{-(\underline{h}(x,r)+\delta)n}$$

infinitely often, which implies that

$$\mu(D_T(x,n,r)) > \mathrm{e}^{-(j+2)\delta n},$$

and so by (3.16) we see that the second choice (3.18) for $\mathbf{C}_n^{A,\delta}$ is used infinitely often. Thus

$$h_{\mu_x^{\mathscr{E}}}(T,\xi) \leqslant \underline{h}(x,r) + 2\delta + \log 2 \liminf_{n\to\infty} \frac{|\mathbf{s}(x)|}{n}.$$

By definition

$$\begin{aligned} |\mathbf{s}(x)| &= \lceil \log_2 |\xi| \rceil \times \left|\{k \mid 0 \leqslant k \leqslant n, T^k z_\ell \in \partial_{2r}\xi\}\right| \\ &\leqslant \lceil \log_2 |\xi| \rceil \times \left|\{k \mid 0 \leqslant k \leqslant n, T^k x \in \partial_{4r}\xi\}\right| \end{aligned}$$

and so (after dividing by n) we can apply the pointwise ergodic theorem to the function $\mathbb{1}_{\partial_{4r}\xi}$ to give

$$\liminf_{n\to\infty} \frac{|\mathbf{s}(x)|}{n} \leqslant \lceil \log_2 |\xi| \rceil \mu_x^{\mathscr{E}}\left(\partial_{4r}\xi\right).$$

If $x \in Y$ (with Y as in (3.15)), $h_{\mu_x^{\mathscr{E}}}(T) < \infty$, and $\underline{h}(x,r) < A$, this gives

$$h_{\mu_x^{\mathscr{E}}}(T) - \varepsilon \leqslant h_{\mu_x^{\mathscr{E}}}(T,\xi) \leqslant \underline{h}(x,r) + 2\delta + \varepsilon \log 2$$

by (3.12). Since $\underline{h}(x,r)$ is monotone increasing as $r \to 0$, it follows that

$$h_{\mu_x^{\mathscr{E}}}(T) - \varepsilon \leqslant \lim_{r\to 0} \underline{h}(x,r) + 2\delta + \varepsilon \log 2.$$

As the last inequality holds for all $A > 0$ and $\delta > 0$, we deduce that

$$h_{\mu_x^{\mathscr{E}}}(T) - \varepsilon \leqslant \lim_{r\to 0} \underline{h}(x,r) + \varepsilon \log 2$$

for $x \in Y$ and $h_{\mu_x^{\mathscr{E}}}(T) < \infty$. Since by construction $\mu(Y) > 1 - 2\varepsilon$ and $\varepsilon > 0$ is arbitrary, we see that

$$h_{\mu_x^{\mathscr{E}}}(T) \leqslant \lim_{r\to 0} \underline{h}(x,r)$$

whenever $h_{\mu_x^{\mathscr{E}}}(T) < \infty$. The case $h_{\mu_x^{\mathscr{E}}}(T) = \infty$ is handled similarly using (3.13) in place of (3.12). □

Exercises for Section 3.4

Exercise 3.4.1. Prove a two-sided version of Theorem 3.12 concerning the growth rate in the two-sided Bowen balls $D_T^{\pm}(x, n, \varepsilon)$.

Exercise 3.4.2. Prove that the assumption that T is a homeomorphism in Theorem 3.12 can be replaced with the assumption that T is a continuous map.

Exercise 3.4.3. Suppose that (X, d) is a compact metric space, $T : X \to X$ is a continuous map, and $\mu \in \mathscr{M}^T(X)$ is an ergodic invariant Borel probability measure.
(a) Show that

$$h_\mu(T) = \lim_{\varepsilon \to 0} \limsup_{n \to \infty} \frac{1}{n} \log M(n, \varepsilon) = \lim_{\varepsilon \to 0} \liminf_{n \to \infty} \frac{1}{n} \log M(n, \varepsilon),$$

where $M(n, \varepsilon)$ is the minimal number of Bowen (n, ε)-balls that are needed to cover a set of μ-measure greater than $1 - \varepsilon$.
(b) Show that the same result holds if $M(n, \varepsilon)$ is changed to be the minimal number of Bowen (n, ε)-balls that are needed to cover a set of μ-measure greater than $\frac{1}{2}$.
(c) What happens to (a) if we drop the assumption of ergodicity?

Notes to Chapter 3

[18] (Page 87) This result was proved in increasingly general settings by Shannon [184] (in his development of information theory), McMillan [132], Carleson [32] (extending to countable partitions of finite entropy), Breiman [28], Ionescu Tulcea [88], Chung [36] and Parry [159] (removing the assumption of an invariant measure). As used in ergodic theory it is usually referred to as the Shannon–McMillan–Breiman Theorem.
[19] (Page 94) Accessible accounts may be found in monographs of Choe [35], Shields [185], and Weiss [208]; particularly relevant results for the discussion above appear in two papers of Orstein and Weiss [155], [156]. Some aspects of coding and symbolic dynamics are discussed by Lind and Marcus [121].
[20] (Page 101) These are often referred to as Bowen–Dinaburg balls, as they were used by both Bowen [24] and Dinaburg [41] to reformulate the definition of topological entropy in terms of how orbits disperse under iteration (see Chapter 5).