Chapter 4
The Entropy Ergodic Theorem

Abstract The goal of this chapter is to prove an ergodic theorem for sample entropy of finite-alphabet random processes. The result is sometimes called the ergodic theorem of information theory or the *asymptotic equipartition (AEP)* theorem, but it is best known as the Shannon-McMillan-Breiman theorem. It provides a common foundation to many of the results of both ergodic theory and information theory.

4.1 History

Shannon [162] first demonstrated the convergence in probability of sample entropy to the entropy rate for stationary ergodic Markov sources. McMillan [123] proved L^1 convergence for stationary ergodic sources and Breiman [20] [21] proved almost everywhere convergence for stationary and ergodic sources. Billingsley [16] extended the result to stationary nonergodic sources. Jacobs [79] [78] extended it to processes dominated by a stationary measure and hence to two-sided AMS processes. Gray and Kieffer [62] extended it to processes asymptotically dominated by a stationary measure and hence to all AMS processes. The generalizations to AMS processes build on the Billingsley theorem for the stationary mean.

Breiman's and Billingsley's approach requires the martingale convergence theorem and embeds the possibly one-sided stationary process into a two-sided process. Ornstein and Weiss [141] developed a proof for the stationary and ergodic case that does not require any martingale theory and considers only positive time and hence does not require any embedding into two-sided processes. The technique was described for both the ordinary ergodic theorem and the entropy ergodic theorem by Shields [165]. In addition, it uses a form of coding argument that is both more direct and more information theoretic in flavor than the traditional martingale proofs. We here follow the Ornstein and Weiss approach for

R.M. Gray, *Entropy and Information Theory*, DOI 10.1007/978-1-4419-7970-4_4,

the stationary ergodic result. We also use some modifications similar to those of Katznelson and Weiss for the proof of the ergodic theorem. We then generalize the result first to nonergodic processes using the "sandwich" technique of Algoet and Cover [7] and then to AMS processes using a variation on a result of [62].

We next state the theorem to serve as a guide through the various steps. We also prove the result for the simple special case of a Markov source, for which the result follows from the usual ergodic theorem.

We consider a directly given finite-alphabet source $\{X_n\}$ described by a distribution m on the sequence measurable space $(\Omega, \mathcal{B})$. Define as previously $X_k^n = (X_k, X_{k+1}, \cdots, X_{k+n-1})$. The subscript is omitted when it is zero. For any random variable Y defined on the sequence space (such as X_k^n) we define the random variable $m(Y)$ by $m(Y)(x) = m(Y = Y(x))$.

Theorem 4.1. *The Entropy Ergodic Theorem*

Given a finite-alphabet AMS source $\{X_n\}$ with process distribution m and stationary mean $\overline{m}$, let $\{\overline{m}_x; x \in \Omega\}$ be the ergodic decomposition of the stationary mean $\overline{m}$. Then

$$\lim_{n\to\infty} \frac{-\ln m(X^n)}{n} = h; \ m - a.e. \text{ and in } L^1(m), \tag{4.1}$$

where $h(x)$ is the invariant function defined by

$$h(x) = \overline{H}_{\overline{m}_x}(X). \tag{4.2}$$

Furthermore,

$$E_m h = \lim_{n\to\infty} \frac{1}{n} H_m(X^n) = \overline{H}_m(X); \tag{4.3}$$

that is, the entropy rate of an AMS process is given by the limit, and

$$\overline{H}_{\overline{m}}(X) = \overline{H}_m(X). \tag{4.4}$$

Comments: The theorem states that the sample entropy using the AMS measure m converges to the entropy rate of the underlying ergodic component of the stationary mean. Thus, for example, if m is itself stationary and ergodic, then the sample entropy converges to the entropy rate of the process m-a.e. and in $L^1(m)$. The $L^1(m)$ convergence follows immediately from the almost everywhere convergence and the fact that sample entropy is uniformly integrable (Lemma 3.7). L^1 convergence in turn immediately implies the left-hand equality of (4.3). Since the limit exists, it is the entropy rate. The final equality states that the entropy rates of an AMS process and its stationary mean are the same. This result follows from (4.2)-(4.3) by the following argument: We have that $\overline{H}_m(X) = E_m h$ and $\overline{H}_{\overline{m}}(X) = E_{\overline{m}} h$, but h is invariant and hence the two expectations are

equal (see, e.g., Lemma 6.3.1 of [55] or Lemma 7.5 of [58]). Thus we need only prove almost everywhere convergence in (4.1) to prove the theorem.

In this section we limit ourselves to the following special case of the theorem that can be proved using the ordinary ergodic theorem without any new techniques.

Lemma 4.1. *Given a finite-alphabet stationary kth order Markov source* $\{X_n\}$*, then there is an invariant function* h *such that*

$$\lim_{n\to\infty} \frac{-\ln m(X^n)}{n} = h;\ m - \text{a.e. } \textit{and in } L^1(m),$$

where h *is defined by*

$$h(x) = -E_{\overline{m}_x} \ln m(X_k|X^k), \tag{4.5}$$

where $\{\overline{m}_x\}$ *is the ergodic decomposition of the stationary mean* $\overline{m}$*. Furthermore,*

$$h(x) = \overline{H}_{\overline{m}_x}(X) = H_{\overline{m}_x}(X_k|X^k). \tag{4.6}$$

Proof of Lemma: We have that

$$-\frac{1}{n}\ln m(X^n) = -\frac{1}{n}\sum_{i=0}^{n-1} \ln m(X_i|X^i).$$

Since the process is kth order Markov with stationary transition probabilites, for $i > k$ we have that

$$m(X_i|X^i) = m(X_i|X_{i-k}, \cdots, X_{i-1}) = m(X_k|X^k)T^{i-k}.$$

The terms $-\ln m(X_i|X^i)$, $i = 0, 1, \cdots, k-1$ have finite expectation and hence are finite m-a.e. so that the ergodic theorem can be applied to deduce

$$\begin{aligned}
\frac{-\ln m(X^n)(x)}{n} &= -\frac{1}{n}\sum_{i=0}^{k-1} \ln m(X_i|X^i)(x) - \frac{1}{n}\sum_{i=k}^{n-1} \ln m(X_k|X^k)(T^{i-k}x) \\
&= -\frac{1}{n}\sum_{i=0}^{k-1} \ln m(X_i|X^i)(x) - \frac{1}{n}\sum_{i=0}^{n-k-1} \ln m(X_k|X^k)(T^i x) \\
&\underset{n\to\infty}{\to} E_{\overline{m}_x}(-\ln m(X_k|X^k)),
\end{aligned}$$

proving the first statement of the lemma. It follows from the ergodic decomposition of Markov sources (see Lemma 8.6.3 of [55] or Lemma 10.5 of [58]) that with probability 1, $\overline{m}_x(X_k|X^k) = m(X_k|\psi(x), X^k) = m(X_k|X^k)$, where ψ is the ergodic component function. This completes the proof. □

We prove the theorem in three steps: The first step considers stationary and ergodic sources and uses the approach of Ornstein and Weiss [141] (see also Shields [165]). The second step removes the requirement for ergodicity. This result will later be seen to provide an information theoretic interpretation of the ergodic decomposition. The third step extends the result to AMS processes by showing that such processes inherit limiting sample entropies from their stationary mean. The later extension of these results to more general relative entropy and information densities will closely parallel the proofs of the second and third steps for the finite case.

In subsequent chapters the definitions of entropy and information will be generalized and corresponding generalizations of the entropy ergodic theorem will be developed in Chapter 11.

4.2 Stationary Ergodic Sources

This section is devoted to proving the entropy ergodic theorem for the special case of stationary ergodic sources. The result was originally proved by Breiman [20]. The original proof first used the martingale convergence theorem to infer the convergence of conditional probabilities of the form $m(X_0|X_{-1}, X_{-2}, \cdots, X_{-k})$ to $m(X_0|X_{-1}, X_{-2}, \cdots)$. This result was combined with an an extended form of the ergodic theorem stating that if $g_k \to g$ as $k \to \infty$ and if g_k is L^1-dominated ($\sup_k |g_k|$ is in L^1), then $1/n \sum_{k=0}^{n-1} g_k T^k$ has the same limit as $1/n \sum_{k=0}^{n-1} g T^k$. Combining these facts yields that that

$$\frac{1}{n} \ln m(X^n) = \frac{1}{n} \sum_{k=0}^{n-1} \ln m(X_k|X^k) = \frac{1}{n} \sum_{k=0}^{n-1} \ln m(X_0|X_{-k}^k) T^k$$

has the same limit as

$$\frac{1}{n} \sum_{k=0}^{n-1} \ln m(X_0|X_{-1}, X_{-2}, \cdots) T^k$$

which, from the usual ergodic theorem, is the expectation

$$E(\ln m(X_0|\mathbf{X}^-) \equiv E(\ln m(X_0|X_{-1}, X_{-2}, \cdots)).$$

As suggested at the end of the preceeding chapter, this should be minus the conditional entropy $H(X_0|X_{-1}, X_{-2}, \cdots)$ which in turn should be the entropy rate $\overline{H}_X$. This approach has three shortcomings: it requires a result from martingale theory which has not been proved here or in the companion volume [55] or [58], it requires an extended ergodic theo-

rem which has similarly not been proved here, and it requires a more advanced definition of entropy which has not yet been introduced. Another approach is the sandwich proof of Algoet and Cover [7]. They show without using martingale theory or the extended ergodic theorem that $1/n \sum_{i=0}^{n-1} \ln m(X_0|X_{-i}^i)T^i$ is asymptotically sandwiched between the entropy rate of a kth order Markov approximation:

$$\frac{1}{n}\sum_{i=k}^{n-1} \ln m(X_0|X_{-k}^k)T^i \underset{n\to\infty}{\to} E_m[\ln m(X_0|X_{-k}^k)] = -H(X_0|X_{-k}^k)$$

and

$$\frac{1}{n}\sum_{i=k}^{n-1} \ln m(X_0|X_{-1}, X_{-2}, \cdots)T^i \underset{n\to\infty}{\to} E_m[\ln m(X_0|X_1, \cdots)] = -H(X_0|X_{-1}, X_{-2}, \cdots).$$

By showing that these two limits are arbitrarily close as $k \to \infty$, the result is proved. The drawback of this approach for present purposes is that again the more advanced notion of conditional entropy given the infinite past is required. Algoet and Cover's proof that the above two entropies are asymptotically close involves martingale theory, but this can be avoided by using Corollary 7.4 as will be seen.

The result can, however, be proved without martingale theory, the extended ergodic theorem, or advanced notions of entropy using the approach of Ornstein and Weiss [141], which is the approach we shall take in this chapter. In a later chapter when the entropy ergodic theorem is generalized to nonfinite alphabets and the convergence of entropy and information densities is proved, the sandwich approach will be used since the appropriate general definitions of entropy will have been developed and the necessary side results will have been proved.

Lemma 4.2. *Given a finite-alphabet source $\{X_n\}$ with a stationary ergodic distribution m, we have that*

$$\lim_{n\to\infty} \frac{-\ln m(X^n)}{n} = h;\ m - \text{a.e.},$$

where $h(x)$ is the invariant function defined by

$$h(x) = \overline{H}_m(X).$$

Proof: Define

$$h_n(x) = -\ln m(X^n)(x) = -\ln m(x^n)$$

and

$$\underline{h}(x) = \liminf_{n\to\infty} \frac{1}{n} h_n(x) = \liminf_{n\to\infty} \frac{-\ln m(x^n)}{n}.$$

Since $m((x_0, \cdots, x_{n-1})) \le m((x_1, \cdots, x_{n-1}))$, we have that

$$h_n(x) \ge h_{n-1}(Tx).$$

Dividing by n and taking the limit infimum of both sides shows that $\underline{h}(x) \ge \underline{h}(Tx)$. Since the $n^{-1}h_n$ are nonnegative and uniformly integrable (Lemma 3.7), we can use Fatou's lemma to deduce that $\underline{h}$ and hence also $\underline{h}T$ are integrable with respect to m. Integrating with respect to the stationary measure m yields

$$\int dm(x)\underline{h}(x) = \int dm(x)\underline{h}(Tx)$$

which can only be true if

$$\underline{h}(x) = \underline{h}(Tx); m - \text{a.e.},$$

that is, if $\underline{h}$ is an invariant function with m-probability one. If $\underline{h}$ is invariant almost everywhere, however, it must be a constant with probability one since m is ergodic (Lemma 6.7.1 of [55] or Lemma 7.12 of [58]). Since it has a finite integral (bounded by $\overline{H}_m(X)$), $\underline{h}$ must also be finite. Henceforth we consider $\underline{h}$ to be a finite constant.

We now proceed with steps that resemble those of the proof of the ergodic theorem in Section 7.2 of [55] or Section 8.1 or [58]. Fix $\epsilon > 0$. We also choose for later use a $\delta > 0$ small enough to have the following properties: If A is the alphabet of X_0 and $\|A\|$ is the finite cardinality of the alphabet, then

$$\delta \ln \|A\| < \epsilon, \tag{4.7}$$

and

$$-\delta \ln \delta - (1-\delta)\ln(1-\delta) \equiv h_2(\delta) < \epsilon. \tag{4.8}$$

The latter property is possible since $h_2(\delta) \to 0$ as $\delta \to 0$.

Define the random variable $n(x)$ to be the smallest integer n for which $n^{-1}h_n(x) \le \underline{h} + \epsilon$. As in the proof of the ergodic theorem, $n(x)$ in general will be large in order to well approximate the limit infimum, but by definition of the limit infimum there must be infinitely many n for which the inequality is true and hence $n(x)$ is everywhere finite, but it is not bounded. Still mimicking the proof of the ergodic theorem, define a set of "bad" sequences $B = \{x : n(x) > N\}$ where N is chosen large enough to ensure that $m(B) < \delta/2$. Define a bounded modification of $n(x)$ by

$$\tilde{n}(x) = \begin{cases} n(x) & x \notin B \\ 1 & x \in B \end{cases}$$

so that $\tilde{n}(x) \le N$ for all $x \in B^c$. We now parse the sequence into variable-length blocks. Iteratively define $n_k(x)$ by

$$\begin{aligned} n_0(x) &= 0 \\ n_1(x) &= \tilde{n}(x) \\ n_2(x) &= n_1(x) + \tilde{n}(T^{n_1(x)}x) = n_1(x) + l_1(x) \\ &\vdots \\ n_{k+1}(x) &= n_k(x) + \tilde{n}(T^{n_k(x)}x) = n_k(x) + l_k(x), \end{aligned}$$

where $l_k(x)$ is the length of the kth block:

$$l_k(x) = \tilde{n}(T^{n_k(x)}x).$$

We have parsed a very long sequence $x^L = (x_0, \cdots, x_{L-1})$, where $L \gg N$, into long blocks $x_{n_k(x)}, \cdots, x_{n_{k+1}(x)-1} = x_{n_k(x)}^{l_k(x)}$ which begin at time $n_k(x)$ and have length $l_k(x)$ for $k = 0, 1, \cdots$. We refer to this parsing as the *block decomposition* of a sequence. The kth block, which begins at time $n_k(x)$, must either have sample entropy satisfying

$$\frac{-\ln m(x_{n_k(x)}^{l_k(x)})}{l_k(x)} \le \underline{h} + \epsilon \tag{4.9}$$

or, equivalently, probability at least

$$m(x_{n_k(x)}^{l_k(x)}) \ge e^{-l_k(x)(\underline{h}+\epsilon)}, \tag{4.10}$$

or it must consist of only a single symbol. Blocks having length 1 ($l_k = 1$) could have the correct sample entropy, that is,

$$\frac{-\ln m(x_{n_k(x)}^{1})}{1} \le \overline{h} + \epsilon,$$

or they could be bad in the sense that they are the first symbol of a sequence with $n > N$; that is,

$$n(T^{n_k(x)}x) > N,$$

or, equivalently,

$$T^{n_k(x)}x \in B.$$

Except for these bad symbols, each of the blocks by construction will have a probability which satisfies the above bound.

Define for nonnegative integers n and positive integers l the sets

$$S(n, l) = \{x : m(X_n^l(x)) \ge e^{-l(\underline{h}+\epsilon)}\},$$

that is, the collection of infinite sequences for which (4.9) and (4.10) hold for a block starting at n and having length l. Observe that for such blocks there cannot be more than $e^{l(\underline{h}+\epsilon)}$ distinct l-tuples for which the

bound holds (lest the probabilities sum to something greater than 1). In symbols this is

$$||S(n,l)|| \le e^{l(\underline{h}+\epsilon)}. \tag{4.11}$$

The ergodic theorem will imply that there cannot be too many single symbol blocks with $n(T^{n_k(x)}x) > N$ because the event has small probability. These facts will be essential to the proof.

Even though we write $\tilde{n}(x)$ as a function of the entire infinite sequence, we can determine its value by observing only the prefix x^N of x since either there is an $n \le N$ for which $n^{-1}\ln m(x^n) \le \underline{h} + \epsilon$ or there is not. Hence there is a function $\hat{n}(x^N)$ such that $\tilde{n}(x) = \hat{n}(x^N)$. Define the finite length sequence event $C = \{x^N : \hat{n}(x^N) = 1 \text{ and } -\ln m(x^1) > \underline{h} + \epsilon\}$, that is, C is the collection of all N-tuples x^N that are prefixes of bad infinite sequences, sequences x for which $n(x) > N$. Thus in particular,

$$x \in B \text{ if and only if } x^N \in C. \tag{4.12}$$

Recall that we parse sequences of length $L \gg N$ and define the set G_L of "good" L-tuples by

$$G_L = \{x^L : \frac{1}{L-N}\sum_{i=0}^{L-N-1} 1_C(x_i^N) \le \delta\},$$

that is, G_L is the collection of all L-tuples which have fewer than $\delta(L-N) \le \delta L$ time slots i for which x_i^N is a prefix of a bad infinite sequence. From (4.12) and the ergodic theorem for stationary ergodic sources we know that m-a.e. we get an x for which

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1} 1_C(x_i^N) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1} 1_B(T^i x) = m(B) \le \frac{\delta}{2}. \tag{4.13}$$

From the definition of a limit, this means that with probability 1 we get an x for which there is an $L_0 = L_0(x)$ such that

$$\frac{1}{L-N}\sum_{i=0}^{L-N-1} 1_C(x_i^N) \le \delta; \text{ for all } L > L_0. \tag{4.14}$$

This follows because if the limit is less than $\delta/2$, there must be an L_0 so large that for larger L the time average is at least no greater than $2\delta/2 = \delta$. We can restate (4.14) as follows: with probability 1 we get an x for which $x^L \in G_L$ for all but a finite number of L. Stating this in negative fashion, we have one of the key properties required by the proof: If $x^L \in G_L$ for all but a finite number of L, then x^L cannot be in the complement G_L^c infinitely often, that is,

$$m(x : x^L \in G_L^c \text{ i.o.}) = 0. \tag{4.15}$$

We now change tack to develop another key result for the proof. For each L we bounded above the cardinality $||G_L||$ of the set of good L-tuples. By construction there are no more than δL bad symbols in an L-tuple in G_L and these can occur in any of at most

$$\sum_{k \le \delta L} \binom{L}{k} \le e^{h_2(\delta)L} \tag{4.16}$$

places, where we have used Lemma 3.6. Eq. (4.16) provides an upper bound on the number of ways that a sequence in G_L can be parsed by the given rules. The bad symbols and the final N symbols in the L-tuple can take on any of the $\|A\|$ different values in the alphabet. Eq. (4.11) bounds the number of finite length sequences that can occur in each of the remaining blocks and hence for any given block decomposition, the number of ways that the remaining blocks can be filled is bounded above by

$$\prod_{k:T^{n_k(x)}x \notin B} e^{l_k(x)(\underline{h}+\epsilon)} = e^{\sum_k l_k(x)(\underline{h}+\epsilon)} = e^{L(\underline{h}+\epsilon)}, \tag{4.17}$$

regardless of the details of the parsing. Combining these bounds we have that

$$||G_L|| \le e^{h_2(\delta)L} \times \|A\|^{\delta L} \times \|A\|^N \times e^{L(\underline{h}+\epsilon)} = e^{h_2(\delta)L + (\delta L + N)\ln\|A\| + L(\underline{h}+\epsilon)}$$

or

$$||G_L|| \le e^{L(\underline{h}+\epsilon+h_2(\delta)+(\delta+\frac{N}{L})\ln\|A\|)}.$$

Since δ satisfies (4.7)–(4.8), we can choose L_1 large enough so that $N \ln\|A\|/L_1 \le \epsilon$ and thereby obtain

$$||G_L|| \le e^{L(\underline{h}+4\epsilon)};\ L \ge L_1. \tag{4.18}$$

This bound provides the second key result in the proof of the lemma. We now combine (4.18) and (4.15) to complete the proof.

Let B_L denote a collection of L-tuples that are bad in the sense of having too large a sample entropy or, equivalently, too small a probability; that is if $x^L \in B_L$, then

$$m(x^L) \le e^{-L(\underline{h}+5\epsilon)}$$

or, equivalently, for any x with prefix x^L

$$h_L(x) \ge \underline{h} + 5\epsilon.$$

The upper bound on $||G_L||$ provides a bound on the probability of $B_L \bigcap G_L$:

$$m(B_L \bigcap G_L) = \sum_{x^L \in B_L \bigcap G_L} m(x^L) \le \sum_{x^L \in G_L} e^{-L(\underline{h}+5\epsilon)}$$

$$\le ||G_L|| e^{-L(\underline{h}+5\epsilon)} \le e^{-\epsilon L}.$$

Recall now that the above bound is true for a fixed $\epsilon > 0$ and for all $L \ge L_1$. Thus

$$\sum_{L=1}^{\infty} m(B_L \bigcap G_L) = \sum_{L=1}^{L_1-1} m(B_L \bigcap G_L) + \sum_{L=L_1}^{\infty} m(B_L \bigcap G_L)$$

$$\le L_1 + \sum_{L=L_1}^{\infty} e^{-\epsilon L} < \infty$$

and hence from the Borel-Cantelli lemma (Lemma 4.6.3 of [55] or Lemma 5.17 of [58]) $m(x : x^L \in B_L \bigcap G_L \text{ i.o.}) = 0$. We also have from (4.15), however, that $m(x : x^L \in G_L^c \text{ i.o. }) = 0$ and hence $x^L \in G_L$ for all but a finite number of L. Thus $x^L \in B_L$ i.o. if and only if $x^L \in B_L \bigcap G_L$ i.o. As this latter event has zero probability, we have shown that $m(x : x^L \in B_L$ i.o.) = 0 and hence

$$\limsup_{L \to \infty} h_L(x) \le \underline{h} + 5\epsilon.$$

Since ϵ is arbitrary we have proved that the limit supremum of the sample entropy $-n^{-1} \ln m(X^n)$ is less than or equal to the limit infimum and therefore that the limit exists and hence with m-probability 1

$$\lim_{n \to \infty} \frac{-\ln m(X^n)}{n} = \underline{h}. \tag{4.19}$$

Since the terms on the left in (4.19) are uniformly integrable from Lemma 3.7, we can integrate to the limit and apply Lemma 3.8 to find that

$$\underline{h} = \lim_{n \to \infty} \int dm(x) \frac{-\ln m(X^n(x))}{n} = \overline{H}_m(X),$$

which completes the proof of the lemma and hence also proves Theorem 4.1 for the special case of stationary ergodic measures. □

4.3 Stationary Nonergodic Sources

Next suppose that a source is stationary with ergodic decomposition $\{m_\lambda; \lambda \in \Lambda\}$ and ergodic component function ψ as in Theorem 1.6. The source will produce with probability one under m an ergodic component m_λ and Lemma 4.2 will hold for this ergodic component. In other words, we should have that

$$\lim_{n\to\infty} -\frac{1}{n}\ln m_\psi(X^n) = \overline{H}_{m_\psi}(X);\ m - \text{a.e.}, \tag{4.20}$$

that is,

$$m(\{x : -\lim_{n\to\infty} \ln m_{\psi(x)}(x^n) = \overline{H}_{m_{\psi(x)}}(X)\}) = 1.$$

This argument is made rigorous in the following lemma.

Lemma 4.3. *Suppose that $\{X_n\}$ is a stationary not necessarily ergodic source with ergodic component function ψ. Then*

$$m(\{x : -\lim_{n\to\infty} \ln m_{\psi(x)}(x^n) = \overline{H}_{m_{\psi(x)}}(X)\}) = 1;\ m - \text{a.e.}. \tag{4.21}$$

Proof: Let

$$G = \{x : -\lim_{n\to\infty} \ln m_{\psi(x)}(x^n) = \overline{H}_{m_{\psi(x)}}(X)\}$$

and let G_λ denote the section of G at λ, that is,

$$G_\lambda = \{x : -\lim_{n\to\infty} \ln m_\lambda(x^n) = \overline{H}_{m_\lambda}(X)\}.$$

From the ergodic decomposition (e.g., Theorem 1.6 or [55], Theorem 8.5.1, [58], Theorem 10.1) and (1.28)

$$m(G) = \int dP_\psi(\lambda) m_\lambda(G),$$

where

$$m_\lambda(G) = m(G|\psi = \lambda) = m(G \bigcap \{x : \psi(x) = \lambda\}|\psi = \lambda)$$

$$= m(G_\lambda|\psi = \lambda) = m_\lambda(G_\lambda)$$

which is 1 for all λ from the stationary ergodic result. Thus

$$m(G) = \int dP_\psi(\lambda) m_\lambda(G_\lambda) = 1.$$

It is straightforward to verify that all of the sets considered are in fact measurable. □

Unfortunately it is not the sample entropy using the distribution of the ergodic component that is of interest, rather it is the original sample entropy for which we wish to prove convergence. The following lemma shows that the two sample entropies converge to the same limit and hence Lemma 4.3 will also provide the limit of the sample entropy with respect to the stationary measure.

Lemma 4.4. *Given a stationary source $\{X_n\}$, let $\{m_\lambda;\ \lambda \in \Lambda\}$ denote the ergodic decomposition and ψ the ergodic component function of Theorem 1.6. Then*

$$\lim_{n\to\infty} \frac{1}{n} \ln \frac{m_\psi(X^n)}{m(X^n)} = 0;\ m - \text{a.e.}$$

Proof: First observe that if $m(a^n)$ is 0, then from the ergodic decomposition with probability 1 $m_\psi(a^n)$ will also be 0. One part is easy. For any $\epsilon > 0$ we have from the Markov inequality that

$$m(\frac{1}{n} \ln \frac{m(X^n)}{m_\psi(X^n)} > \epsilon) = m(\frac{m(X^n)}{m_\psi(X^n)} > e^{n\epsilon}) \le E_m(\frac{m(X^n)}{m_\psi(X^n)}) e^{-n\epsilon}.$$

The expectation, however, can be evaluated as follows: Let $A_n^{(\lambda)} = \{a^n : m_\lambda(a^n) > 0\}$. Then

$$E_m\left(\frac{m(X^n)}{m_\psi(X^n)}\right) = \int dP_\psi(\lambda) \sum_{a^n \in A_n} \frac{m(a^n)}{m_\lambda(a^n)} m_\lambda(a^n) = \int dP_\psi(\lambda) m(A_n^{(\lambda)}) \le 1,$$

where P_ψ is the distribution of ψ. Thus

$$m(\frac{1}{n} \ln \frac{m(X^n)}{m_\psi(X^n)} > \epsilon) \le e^{-n\epsilon}.$$

and hence

$$\sum_{n=1}^{\infty} m(\frac{1}{n} \ln \frac{m(X^n)}{m_\psi(X^n)} > \epsilon) < \infty$$

and hence from the Borel-Cantelli lemma

$$m(\frac{1}{n} \ln \frac{m(X^n)}{m_\psi(X^n)} > \epsilon \text{ i.o.}) = 0$$

and hence with m probability 1

$$\limsup_{n\to\infty} \frac{1}{n} \ln \frac{m(X^n)}{m_\psi(X^n)} \le \epsilon.$$

Since ϵ is arbitrary,

$$\limsup_{n\to\infty} \frac{1}{n} \ln \frac{m(X^n)}{m_\psi(X^n)} \le 0;\ m - \text{a.e.} \tag{4.22}$$

For later use we restate this as

$$\liminf_{n\to\infty} \frac{1}{n} \ln \frac{m_\psi(X^n)}{m(X^n)} \ge 0;\ m - \text{a.e.} \tag{4.23}$$

Now turn to the converse inequality. For any positive integer k, we can construct a stationary k-step Markov approximation to m as in Section 3.7 that is, construct a process $m^{(k)}$ with the conditional probabilities

$$m^{(k)}(X_n \in F|X^n) = m^{(k)}(X_n \in F|X_{n-k}^k) = m(X_n \in F|X_{n-k}^k)$$

and the same kth order distributions $m^{(k)}(X^k \in F) = m(X^k \in F)$. Consider the probability

$$m(\frac{1}{n}\ln\frac{m^{(k)}(X^n)}{m(X^n)} \geq \epsilon) = m(\frac{m^{(k)}(X^n)}{m(X^n)} \geq e^{n\epsilon}) \leq E_m(\frac{m^{(k)}(X^n)}{m(X^n)})e^{-n\epsilon}.$$

The expectation is evaluated as

$$\sum_{x^n}\frac{m^{(k)}(x^n)}{m(x^n)}m(x^n) = 1$$

and hence we again have using Borel-Cantelli that

$$\limsup_{n\to\infty}\frac{1}{n}\ln\frac{m^{(k)}(X^n)}{m(X^n)} \leq 0.$$

Apply the usual ergodic theorem to conclude that with probability 1 under m

$$\limsup_{n\to\infty}\frac{1}{n}\ln\frac{1}{m(X^n)} \leq \lim_{n\to\infty}\frac{1}{n}\ln\frac{1}{m^{(k)}(X^n)} = E_{m_\psi}[-\ln m(X_k|X^k)].$$

Combining this result with (4.20) and Lemma 3.10 yields

$$\limsup_{n\to\infty}\frac{1}{n}\ln\frac{m_\psi(X^n)}{m(X^n)} \leq -\overline{H}_{m_\psi}(X) - E_{m_\psi}[\ln m(X_k|X^k)]. = \overline{H}_{m_\psi||m^{(k)}}(X).$$

This bound holds for any integer k and hence it must also be true that m-a.e. the following holds:

$$\limsup_{n\to\infty}\frac{1}{n}\ln\frac{m_\psi(X^n)}{m(X^n)} \leq \inf_k \overline{H}_{m_\psi||m^{(k)}}(X) \equiv \zeta. \tag{4.24}$$

In order to evaluate ζ we apply the ergodic decomposition of relative entropy rate (Corollary 3.5) and the ordinary ergodic decomposition to write

$$\begin{aligned}\int dP_\psi \zeta &= \int dP_\psi \inf_k \overline{H}_{m_\psi||m^{(k)}}(X)\\ &\leq \inf_k \int dP_\psi \overline{H}_{m_\psi||m^{(k)}}(X) = \inf_k \overline{H}_{m||m^{(k)}}(X).\end{aligned}$$

From Theorem 3.4, the right hand term is 0. If the integral of a nonnegative function is 0, the integrand must itself be 0 with probability one. Thus (4.24) becomes

$$\limsup_{n\to\infty} \frac{1}{n} \ln \frac{m_\psi(X^n)}{m(X^n)} \le 0,$$

which with (4.23) completes the proof of the lemma. □

We shall later see that the quantity

$$i_n(X^n;\psi) = \frac{1}{n} \ln \frac{m_\psi(X^n)}{m(X^n)}$$

is the sample mutual information (in a generalized sense so that it applies to the usually non-discrete ψ) and hence the lemma states that the normalized sample mutual information between the process outputs and the ergodic component function goes to 0 as the number of samples goes to infinity.

The two previous lemmas immediately yield the following result.

Corollary 4.1. *The conclusions of Theorem 4.1 hold for sources that are stationary.*

4.4 AMS Sources

The principal idea required to extend the entropy theorem from stationary sources to AMS sources is contained in Lemma 4.6. It shows that an AMS source inherits sample entropy properties from an asymptotically dominating stationary source (just as it inherits ordinary ergodic properties from such a source). The result is originally due to Gray and Kieffer [62], but the proof here is somewhat different. The tough part here is handling the fact that the sample average being considered depends on a specific measure. From Theorem 1.2, the stationary mean of an AMS source dominates the original source on tail events, that is, events in $\mathcal{F}_\infty$. We begin by showing that certain important events can be recast as tail events, that is, they can be determined by looking at only samples in the arbitrarily distant future. The following result is of this variety: It implies that sample entropy is unaffected by the starting time.

Lemma 4.5. *Let $\{X_n\}$ be a finite-alphabet source with distribution m. Recall that $X_k^n = (X_k, X_{k+1}, \cdots, X_{k+n-1})$ and define the information density*

$$i(X^k; X_k^{n-k}) = \ln \frac{m(X^n)}{m(X^k)m(X_k^{n-k})}.$$

Then

$$\lim_{n\to\infty} \frac{1}{n} i(X^k; X_k^{n-k}) = 0; \ m - \text{a.e.}$$

Comment: The lemma states that with probability 1 the per-sample mutual information density between the first k samples and future samples goes to zero in the limit. Equivalently, limits of $n^{-1} \ln m(X^n)$ will be the same as limits of $n^{-1} \ln m(X_k^{n-k})$ for any finite k. Note that the result does not require even that the source be AMS. The lemma is a direct consequence of Lemma 3.19.
Proof: Define the distribution $p = m_{X^k} \times m_{X_k, X_{k+1}, \dots}$, that is, a distribution for which all samples after the first k are independent of the first k samples. Thus, in particular, $p(X^n) = m(X^k) m(X_k^n)$. We will show that $p \gg m$, in which case the lemma will follow from Lemma 3.19. Suppose that $p(F) = 0$. If we denote $X_k^+ = X_k, X_{k+1}, \cdots$, then

$$0 = p(F) = \sum_{x^k} m(x^k) m_{X_k^+}(F_{x^k}),$$

where F_{x^k} is the section $\{x_k^+ : (x^k, x_k^+) = x \in F\}$. For the above relation to hold, we must have $m_{X_k^+}(F_{x^k}) = 0$ for all x^k with $m(x^k) \neq 0$. We also have, however, that

$$m(F) = \sum_{a^k} m(X^k = a^k, X_k^+ \in F_{a^k})$$

$$= \sum_{a^k} m(X^k = a^k | X_k^+ \in F_{a^k}) m(X_k^+ \in F_{a^k}).$$

But this sum must be 0 since the rightmost terms are 0 for all a^k for which $m(X^k = a^k)$ is not 0. (Observe that we must have $m(X^k = a^k | X_k^+ \in F_{a^k}) = 0$ if $m(X_k^+ \in F_{a^k}) \neq 0$ since otherwise $m(X^k = a^k) \geq m(X^k = a^k, X_k^+ \in F_{a^k}) > 0$, yielding a contradiction.) Thus $p \gg m$ and the lemma is proved. □

For later use we note that we have shown that a joint distribution is dominated by a product of its marginals if one of the marginal distributions is discrete.

Lemma 4.6. *Suppose that $\{X_n\}$ is an AMS source with distribution m and suppose that $\overline{m}$ is a stationary source that asymptotically dominates m (e.g., $\overline{m}$ is the stationary mean). If there is an invariant function h such that*

$$\lim_{n \to \infty} -\frac{1}{n} \ln \overline{m}(X^n) = h; \ \overline{m} - \text{a.e.},$$

then also,

$$\lim_{n \to \infty} -\frac{1}{n} \ln m(X^n) = h; \ m - \text{a.e.}$$

Proof: For any k we can write using the chain rule for densities

$$-\frac{1}{n}\ln m(X^n) + \frac{1}{n}\ln m(X_k^{n-k}) = -\frac{1}{n}\ln m(X^k|X_k^{n-k})$$
$$= -\frac{1}{n}i(X^k;X_k^{n-k}) - \frac{1}{n}\ln m(X^k).$$

From the previous lemma and from the fact that $H_m(X^k) = -E_m \ln m(X^k)$ is finite, the right hand terms converge to 0 as $n \to \infty$ and hence for any k

$$\lim_{n\to\infty} -\frac{1}{n}\ln m(X^k|X_k^{n-k}) =$$
$$\lim_{n\to\infty}(-\frac{1}{n}\ln m(X^n) + \frac{1}{n}\ln m(X_k^{n-k})) = 0;\ m-\text{a.e.} \quad (4.25)$$

This implies that there is a subsequence $k(n) \to \infty$ such that

$$-\frac{1}{n}\ln m(X^{k(n)}|X_{k(n)}^{n-k(n)}) = -\frac{1}{n}\ln m(X^n) - \frac{1}{n}\ln m(X_{k(n)}^{n-k}(n)) \to 0;\ m-\text{a.e.} \quad (4.26)$$

To see this, observe that (4.25) ensures that for each k there is an $N(k)$ large enough so that $N(k) > N(k-1)$ and

$$m(|-\frac{1}{N(k)}\ln m(X^k|X_k^{N(k)-k})| > 2^{-k}) \le 2^{-k}. \quad (4.27)$$

Applying the Borel-Cantelli lemma implies that for any ϵ,

$$m(|-1/N(k)\ln m(X^k|X_k^{N(k)-k})| > \epsilon \text{ i.o.}) = 0.$$

Now let $k(n) = k$ for $N(k) \le n < N(k+1)$. Then

$$m(|-1/n\ln m(X^{k(n)}|X_{k(n)}^{n-k(n)})| > \epsilon \text{ i.o.}) = 0$$

and therefore

$$\lim_{n\to\infty}\left(-\frac{1}{n}\ln m(X^n) + \frac{1}{n}\ln m(X_{k(n)}^{n-k(n)})\right) = 0;\ m-\text{a.e.}$$

as claimed in (4.26).

In a similar manner we can also choose the sequence so that

$$\lim_{n\to\infty}\left(-\frac{1}{n}\ln \overline{m}(X^n) + \frac{1}{n}\ln \overline{m}(X_{k(n)}^{n-k(n)})\right) = 0;\ \overline{m}-\text{a.e.},$$

that is, we can choose $N(k)$ so that (4.27) simultaneously holds for both m and $\overline{m}$. Invoking the entropy ergodic theorem for the stationary $\overline{m}$ (Corollary 4.3) we have therefore that

$$\lim_{n\to\infty} -\frac{1}{n}\ln \overline{m}(X_{k(n)}^{n-k(n)}) = \overline{h};\ \overline{m}-\text{a.e.}. \quad (4.28)$$

From Markov's inequality (Lemma 4.4.3 of [55] or Lemma 5.8 of [58])

$$\begin{aligned}\overline{m}(-\frac{1}{n}\ln m(X_k^n) \le -\frac{1}{n}\ln \overline{m}(X_k^n) - \epsilon) &= \overline{m}(\frac{m(X_k^n)}{\overline{m}(X_k^n)} \ge e^{n\epsilon}) \\ &\le e^{-n\epsilon} E_{\overline{m}} \frac{m(X_k^{n-k})}{\overline{m}(X_k^{n-k})} \\ &= e^{-n\epsilon} \sum_{x_k^{n-k}:\overline{m}(x_k^{n-k})\neq 0} \frac{m(x_k^{n-k})}{\overline{m}(x_k^{n-k})} \overline{m}(x_k^{n-k}) \\ &\le e^{-n\epsilon}.\end{aligned}$$

Hence taking $k = k(n)$ and again invoking the Borel-Cantelli lemma we have that

$$\overline{m}(-\frac{1}{n}\ln m(X_{k(n)}^{n-k(n)}) \le -\frac{1}{n}\ln \overline{m}(X_{k(n)}^{n-k(n)}) - \epsilon \text{ i.o.}) = 0$$

or, equivalently, that

$$\liminf_{n\to\infty} -\frac{1}{n}\ln \frac{m(X_{k(n)}^{n-k(n)})}{\overline{m}(X_{k(n)}^{n-k(n)})} \ge 0;\ \overline{m}\text{ – a.e.} \tag{4.29}$$

Therefore from (4.28)

$$\liminf_{n\to\infty} -\frac{1}{n}\ln m(X_{k(n)}^{n-k(n)}) \ge h;\ \overline{m}\text{ – a.e..} \tag{4.30}$$

The above event is in the tail σ-field $\mathcal{F}_\infty = \bigcap_n \sigma(X_n, X_{n+1}, \cdots)$ since it can be determined from $X_{k(n)}, \cdots$ for arbitrarily large n and since h is invariant. Since $\overline{m}$ dominates m on the tail σ-field (Theorem 1.3), we have also

$$\liminf_{n\to\infty} -\frac{1}{n}\ln m(X_{k(n)}^{n-k(n)}) \ge h;\ m\text{ – a.e.}$$

and hence by (4.26)

$$\liminf_{n\to\infty} -\frac{1}{n}\ln m(X^n) \ge h;\ m\text{ – a.e.}$$

which proves half of the lemma. Since

$$\overline{m}(\lim_{n\to\infty} -\frac{1}{n}\ln \overline{m}(X^n) \ne h) = 0$$

and since $\overline{m}$ asymptotically dominates m (Theorem 1.2), given $\epsilon > 0$ there is a k such that

$$m(\lim_{n\to\infty} -\frac{1}{n}\ln \overline{m}(X_k^n) = h) \ge 1 - \epsilon.$$

Again applying Markov's inequality and the Borel-Cantelli lemma as in the development of (4.28) we have that

$$\liminf_{n\to\infty} -\frac{1}{n}\ln\frac{\overline{m}(X_k^n)}{m(X_k^n)} \geq 0;\ m-\text{a.e},$$

which implies that

$$m(\limsup_{n\to\infty} -\frac{1}{n}\ln m(X_k^n) \leq h) \geq 1-\epsilon$$

and hence also that

$$m(\limsup_{n\to\infty} -\frac{1}{n}\ln m(X^n) \leq h) \geq 1-\epsilon.$$

Since ϵ can be made arbitrarily small, this proves that m-a.e.

$$\limsup_{n\to\infty} -n^{-1}\ln m(X^n) \leq h,$$

which completes the proof of the lemma. □

The lemma combined with Corollary 4.3 completes the proof of Theorem 4.1. □

Theorem 4.1 and Lemma 2.5 immediately yield the following corollary stating that a stationary coding of an AMS process has a well defined entropy rate given by a limit, as in the case of a stationary process.

Corollary 4.2. *Theorem 4.1 If f is a stationary coding of an AMS process, then*

$$\overline{H}(f) = \lim_{n\to\infty}\frac{1}{n}H(f^n).$$

4.5 The Asymptotic Equipartition Property

Since convergence almost everywhere implies convergence in probability, Theorem 4.1 has the following implication: Suppose that $\{X_n\}$ is an AMS ergodic source with entropy rate $\overline{H}$. Given $\epsilon > 0$ there is an N such that for all $n > N$ the set

$$G_n = \{x^n : |n^{-1}h_n(x) - \overline{H}| \geq \epsilon\} = \{x^n : e^{-n(\overline{H}+\epsilon)} \leq m(x^n) \leq e^{-n(\overline{H}-\epsilon)}\}$$

has probability greater then $1-\epsilon$. Furthermore, as in the proof of the theorem, there can be no more than $e^{n(\overline{H}+\epsilon)}$ n-tuples in G_n. Thus there are two sets of n-tuples: a "good" set of approximately $e^{n\overline{H}}$ n-tuples having approximately equal probability of $e^{-n\overline{H}}$ and the complement of this set which has small total probability. The set of good sequences are

often referred to as "typical sequences" or "entropy-typical sequences" in the information theory literature and in this form the theorem is called the asymptotic equipartition property or the AEP.

As a first information theoretic application of an ergodic theorem, we consider a simple coding scheme called an "almost noiseless" or "almost lossless" source code. As we often do, we consider logarithms to the base 2 when considering specific coding applications. Suppose that a random process $\{X_n\}$ has a finite alphabet A with cardinality $\|A\|$ and entropy rate $\overline{H}$. Suppose that $\overline{H} < \log \|A\|$, e.g., A might have 16 symbols, but the entropy rate is slightly less than 2 bits per symbol rather than $\log 16 = 4$. Larger alphabets cost money in either storage or communication applications. For example, to communicate a source with a 16 letter alphabet sending one letter per second without using any coding and using a binary communication system we would need to send 4 binary symbols (or four *bits*) for each source letter and hence 4 bits per second would be required. If the alphabet only had 4 letters, we would need to send only 2 bits per second. The question is the following: Since our source has an alphabet of size 16 but an entropy rate of less than 2, can we code the original source into a new source with an alphabet of only $4 = 2^2$ letters so as to communicate the source at the smaller rate and yet have the receiver be able to recover the original source? The AEP suggests a technique for accomplishing this provided we are willing to tolerate rare errors.

We construct a block code of the original source by first picking a small ϵ and a δ small enough so that $\overline{H} + \delta < 2$. Choose a large enough n so that the AEP holds giving a set G_n of good sequences as above with probability greater than $1 - \epsilon$. Index this collection of fewer than $2^{n(\overline{H}+\delta)} < 2^{2n}$ sequences using binary $2n$-tuples. The source X_k is parsed into blocks of length n as $X_{kn}^n = (X_{kn}, X_{kn+1}, \cdots, X_{(k+1)n})$ and each block is encoded into a binary $2n$-tuple as follows: If the source n-tuple is in G_n, the codeword is its binary $2n$-tuple index. Select one of the unused binary $2n$-tuples as the error index and whenever an n-tuple is not in G_n, the error index is the codeword. The receiver or decoder than uses the received index and decodes it as the appropriate n-tuple in G_n. If the error index is received, the decoder can declare an arbitrary source sequence or just declare an error. With probability at least $1 - \epsilon$ a source n-tuple at a particular time will be in G_n and hence it will be correctly decoded. We can make this probability as small as desired by taking n large enough, but we cannot in general make it 0.

The above simple scheme is an example of a block coding scheme as considered in Section 2.7. If considered as a mapping from sequences into sequences, the map is not stationary, but it is block stationary in the sense that shifting an input block by n results in a corresponding block shift of the encoded sequence by $2n$ binary symbols.