The Annals of Probability
1988, Vol. 16, No. 2, 899–909

A SANDWICH PROOF OF THE SHANNON–McMILLAN–BREIMAN THEOREM

BY PAUL H. ALGOET[1] AND THOMAS M. COVER[2]

Boston University and Stanford University

Let $\{X_t\}$ be a stationary ergodic process with distribution P admitting densities $p(x_0, \ldots, x_{n-1})$ relative to a reference measure M that is finite order Markov with stationary transition kernel. Let $I_M(P)$ denote the relative entropy rate. Then

$$n^{-1}\log p(X_0, \ldots, X_{n-1}) \to I_M(P) \quad \text{a.s. } (P).$$

We present an elementary proof of the Shannon–McMillan–Breiman theorem and the preceding generalization, obviating the need to verify integrability conditions and also covering the case $I_M(P) = \infty$. A sandwich argument reduces the proof to direct applications of the ergodic theorem.

1. Introduction. If $p(x_0, \ldots, x_{n-1})$ and $p(x_n|x_{n-1}, \ldots, x_0)$ denote joint and conditional probability mass functions of a stationary ergodic process $\{X_t\}$ taking values in a countable set $\mathcal{X}$, then the Shannon–McMillan–Breiman theorem asserts that

$$(1) \quad -n^{-1}\log p(X_0, \ldots, X_{n-1}) = -n^{-1}\sum_{t=0}^{n-1} \log p(X_t|X_{t-1}, \ldots, X_0) \to H \quad \text{a.s.},$$

where $H = \lim_k \downarrow E\{-\log p(X_k|X_{k-1}, \ldots, X_0)\}$ is the entropy rate of $\{X_t\}$. This individual ergodic theorem of information theory was proved first by Breiman (1957/1960) for finite $\mathcal{X}$, and later by Chung (1961, 1962) for countably infinite $\mathcal{X}$ under the hypothesis $E\{-\log p(X_0)\} < \infty$. Convergence in probability already implies the existence of a set of roughly (to first order in the exponent) $\exp(nH)$ typical sequences of length n all having roughly equal probability $\exp(-nH)$; cf. Shannon (1948). McMillan (1953) called this the asymptotic equipartition property or AEP, and proved convergence in L^1.

The AEP has recently been generalized to processes with densities. Indeed, suppose $\mathcal{X}$ is a standard Borel space and $(\Omega, \mathcal{F})$ designates the sequence space $\mathcal{X}_0^\infty = \prod_{t=0}^\infty \mathcal{X}$ with its Borel σ-field. Let P denote the distribution of a stationary ergodic process $\{X_t\}$ on $(\Omega, \mathcal{F})$, and M a finite order Markov distribution with stationary transition kernel. We assume absolute continuity of the nth order marginal of P with respect to the nth order marginal of M and denote the corresponding density by $p(x_0, \ldots, x_{n-1})$, for n finite. The generalized

Received August 1985; revised April 1987.

[1]Partially supported by National Science Foundation Grant ECS-82-11568 and Joint Services Electronics Program Grant DAAG 29-84-K-0047.

[2]Partially supported by National Science Foundation Grant ECS-82-11568 and by Bell Communications Research.

AMS 1980 *subject classifications.* Primary 28D05, 94A17; secondary 28A65, 28D20, 60F15.

Key words and phrases. Shannon–McMillan–Breiman theorem, asymptotic equipartition property (AEP), ergodic theorem of information theory, relative entropy rate, likelihood ratio, sandwich argument, Markov approximation, asymptotically mean stationary.

Shannon–McMillan–Breiman theorem then states that

$$(2) \qquad n^{-1}\log p(X_0,\ldots,X_{n-1}) \to I_M(P) \quad \text{a.s. } (P),$$

where $I_M(P) = \lim_k \uparrow E\{\log p(X_k|X_{k-1},\ldots,X_0)\}$ is the relative entropy rate of the true distribution P with respect to the reference measure M. Thus the likelihood ratio $p(X_0,\ldots,X_{n-1})$ will grow exponentially fast almost surely with limiting rate $I_M(P)$. The outcomes $X_0,\ldots,X_{n-1}$ will be distributed with nearly uniform density $\exp(nI_M(P))$ over the typical set $\{|n^{-1}\log p(X_0,\ldots,X_{n-1}) - I_M(P)| < \varepsilon\}$ whose M-measure is roughly $\exp(-nI_M(P))$, smallest possible for sets whose P-measure is bounded away from 0. Stein's lemma [cf. Chernoff (1956) for the i.i.d. case] identifies $I_M(P)$ as the best exponential decay rate of the error probability when discriminating the null hypothesis P against the alternative M on the basis of a growing number of observations $X_0, X_1,\ldots,X_{n-1}$.

This strong AEP was recently proved by Barron (1985) and Orey (1985), after Moy (1960, 1961), Perez (1964, 1974), and Kieffer (1973/1976, 1974) suggested its validity and proved L^1 convergence. Barron and Orey invoke Breiman's (1957/1960) extended ergodic theorem when they observe that $g_t(\omega) = \log p(X_0|X_{-1},\ldots,X_{-t})$ almost surely converges to $g(\omega) = \log p(X_0|X_{-1}, X_{-2},\ldots)$ and hence (writing T for the usual shift on $\mathscr{X}_{-\infty}^{\infty} = \prod_{t=-\infty}^{\infty}\mathscr{X}$),

$$(3) \quad n^{-1}\log p(X_0,\ldots,X_{n-1}) = n^{-1}\sum_{t=0}^{n-1} g_t(T^t\omega) \to E\{g\} = I_M(P) \quad \text{a.s. } (P)$$

provided $\{g_t - g\}_{t=k}^{\infty}$ is dominated in $L^1(P)$ for some finite k. The essential contribution of these authors was proving that $E\{\sup_{t=k}^{\infty}|g - g_t|\} < \infty$. We shall present an elementary proof of the AEP which obviates the need to verify this integrability condition. We argue that $p(X_0,\ldots,X_{n-1})$ is sandwiched in asymptotic growth rate between the kth order Markov approximation $p^k(X_0,\ldots,X_{n-1})$ and the infinite order approximation $p(X_0,\ldots,X_{n-1}|X_{-1},\ldots)$, with no gap as $k \to \infty$.

The paper is organized as follows. We exhibit the essence of the sandwich argument in Section 2, while proving Breiman's AEP for a finite-valued stationary ergodic process. The generalized AEP is proved in Section 3 for processes with values in a standard Borel space. The reference measure M is finite order Markov with stationary transition kernel, and we must define the densities and consider the possibility of an infinite relative entropy rate as limiting expectation. In Section 4 we prove the AEP for processes that are stationary but not necessarily ergodic, and for asymptotically mean stationary processes satisfying an extra hypothesis.

If $i \le j$ are finite indices then X_i^j, $X_{-\infty}^{i-1}$ and X_{j+1}^{∞} will denote sequences in the product spaces $\mathscr{X}_i^j = \prod_{t=i}^{j}\mathscr{X}$, $\mathscr{X}_{-\infty}^{i-1} = \prod_{t=\infty}^{i-1}\mathscr{X}$ and $\mathscr{X}_{j+1}^{\infty} = \prod_{t=j+1}^{\infty}\mathscr{X}$.

2. A sandwich proof of the AEP for finite-valued random processes. We prove the AEP for a stationary ergodic process $\{X_t\}$ with values in a finite set $\mathscr{X}$. The *entropy rate* H of $\{X_t\}$ is defined as $\lim_k \downarrow H^k$ where

$$(4) \quad H^k = E\{-\log p(X_k|X_{k-1},\ldots,X_0)\} = E\{-\log p(X_0|X_{-1},\ldots,X_{-k})\}.$$

$H^k = E\{-\log p(X_k|X_0^{k-1})\}$ is equal to $E\{-\log p(X_0|X_{-k}^{-1})\}$ by stationarity, and H^k is nonincreasing by Jensen's inequality. H can also be defined as the Cesàro limit,

$$H = \lim_n \downarrow n^{-1}E\{-\log p(X_0,\ldots,X_{n-1})\} = \lim_n \downarrow n^{-1}\sum_{t=0}^{n-1} H^k. \tag{5}$$

It will be crucial [and argued in the following; see (17)–(18)] that $H^k \searrow H = H^\infty$, where

$$H^\infty = E\{-\log p(X_0|X_{-1}, X_{-2},\ldots)\}. \tag{6}$$

The following lemma will be used when g_n is the likelihood ratio of an alternative measure relative to the true distribution of $\{X_t\}$.

LEMMA 1. *If $\{g_n\}$ is a sequence of positive random variables such that $E\{g_n\} \le 1$ for all n, then*

$$\limsup_n n^{-1}\log g_n \le 0 \quad a.s. \tag{7}$$

PROOF. If $\varepsilon > 0$, then $P\{n^{-1}\log g_n \ge \varepsilon\} = P\{g_n \ge \exp(n\varepsilon)\} \le \exp(-n\varepsilon)$ by Markov's inequality. But $\Sigma_n \exp(-n\varepsilon) < \infty$ and hence $\limsup_n n^{-1}\log g_n \le \varepsilon$ a.s. by the Borel–Cantelli lemma. The lemma follows since $\varepsilon > 0$ was arbitrary. □

THEOREM 1 (Breiman's AEP). *If H is the entropy rate of a finite-valued stationary ergodic process $\{X_t\}$, then*

$$\begin{aligned} &-n^{-1}\log p(X_0,\ldots,X_{n-1}) \\ &\quad = -n^{-1}\sum_{t=0}^{n-1}\log p(X_t|X_{t-1},\ldots,X_0) \to H \quad a.s. \end{aligned} \tag{8}$$

PROOF. We argue that the likelihood growth rate $-n^{-1}\log p(X_0,\ldots,X_{n-1})$ is asymptotically sandwiched between the upper bound H^k and the lower bound H^∞, for all $k \ge 0$. The AEP will follow since the sandwich closes in the limit as $k \to \infty$.

The kth order Markov approximation of the probability $p(X_0,\ldots,X_{n-1})$ is defined for large n ($n \ge k$) as

$$p^k(X_0,\ldots,X_{n-1}) = p(X_0,\ldots,X_{k-1})\prod_{t=k}^{n-1} p(X_t|X_{t-1},\ldots,X_{t-k}). \tag{9}$$

In view of the expansions

$$\begin{aligned} &-n^{-1}\log p^k(X_0,\ldots,X_{n-1}) \\ &\quad = -n^{-1}\log p(X_0,\ldots,X_{k-1}) - n^{-1}\sum_{t=k}^{n-1}\log p(X_t|X_{t-1},\ldots,X_{t-k}) \end{aligned} \tag{10}$$

and

$$
\begin{aligned}
&-n^{-1}\log p(X_0,\ldots,X_{n-1}|X_{-1},\ldots)\\
&\quad=-n^{-1}\sum_{t=0}^{n-1}\log p(X_t|X_{t-1},\ldots,X_0,X_{-1},\ldots),
\end{aligned}
\tag{11}
$$

the ergodic theorem asserts that

$$
-n^{-1}\log p^k(X_0,\ldots,X_{n-1})\to H^k=E\{-\log p(X_k|X_{k-1},\ldots,X_0)\}\quad\text{a.s.}
\tag{12}
$$

and

$$
\begin{aligned}
&-n^{-1}\log p(X_0,\ldots,X_{n-1}|X_{-1},\ldots)\\
&\quad\to H^\infty=E\{-\log p(X_0|X_{-1},\ldots)\}\quad\text{a.s.}
\end{aligned}
\tag{13}
$$

The expectation of the likelihood ratio of an alternative measure relative to the true distribution is equal to the mass of the absolutely continuous part of the alternative measure, and this is no larger than its total mass. Thus

$$
E\left\{\frac{p^k(X_0,\ldots,X_{n-1})}{p(X_0,\ldots,X_{n-1})}\right\}\le 1
\tag{14}
$$

and

$$
E\left\{\frac{p(X_0,\ldots,X_{n-1})}{p(X_0,\ldots,X_{n-1}|X_{-1},X_{-2},\ldots)}\right\}\le 1.
\tag{15}
$$

By Markov's inequality and the Borel–Cantelli lemma (cf. proof of Lemma 1),

$$
\limsup_n\, n^{-1}\log\left(\frac{p^k(X_0,\ldots,X_{n-1})}{p(X_0,\ldots,X_{n-1})}\right)\le 0\quad\text{a.s.}
\tag{16}
$$

and

$$
\limsup_n\, n^{-1}\log\left(\frac{p(X_0,\ldots,X_{n-1})}{p(X_0,\ldots,X_{n-1}|X_{-1},X_{-2},\ldots)}\right)\le 0\quad\text{a.s.}
\tag{17}
$$

Writing the log-likelihood ratios in (16) and (17) as differences of log-likelihoods and applying the limit theorems (10) and (13) we obtain the chain of asymptotic inequalities

$$
\begin{aligned}
H^k&=E\{-\log p(X_0|X_{-1},\ldots,X_{-k})\}\\
&\ge\limsup_n\,-n^{-1}\log p(X_0,\ldots,X_{n-1})\\
&\ge\liminf_n\,-n^{-1}\log p(X_0,\ldots,X_{n-1})\\
&\ge E\{-\log p(X_0|X_{-1},X_{-2},\ldots)\}=H^\infty\quad\text{a.s.}
\end{aligned}
\tag{18}
$$

It remains to show that no gap exists between H^k and H^∞ in the limit as $k\to\infty$. Indeed, Lévy's martingale convergence theorem for conditional probabil-

ities asserts that

$$p(x_0|X_{-k}^{-1}) \to p(x_0|X_{-\infty}^{-1}) \quad \text{a.s. for all } x_0 \in \mathscr{X}. \tag{19}$$

Since $\mathscr{X}$ is finite and $p \log p$ is bounded and continuous in p as $0 \le p \le 1$, the bounded convergence theorem implies that

$$\begin{aligned} H^k &= E\Big\{-\sum_{x_0} p(x_0|X_{-k}^{-1}) \log p(x_0|X_{-k}^{-1})\Big\} \\ &\to E\Big\{-\sum_{x_0} p(x_0|X_{-\infty}^{-1}) \log p(x_0|X_{-\infty}^{-1})\Big\} = H^\infty. \end{aligned} \tag{20}$$

Thus $H^k \searrow H = H^\infty$ as $k \to \infty$, and the AEP follows. □

3. The generalized AEP for processes with densities. Let $\mathscr{X}$ designate a standard Borel space, $(\Omega, \mathscr{F})$ the sequence space $\mathscr{X}_0^\infty$ with its Borel σ-field, T the left shift on $(\Omega, \mathscr{F})$, and $X_t(\omega) = X_0(T^t\omega)$ the usual coordinate projections. If P and Q are probability distributions on $(\Omega, \mathscr{F})$ and $\mathscr{G}$ is a sub-σ-field of $\mathscr{F}$, then $(dQ/dP)|_{\mathscr{G}}$ will denote the likelihood ratio of the restriction $Q|_{\mathscr{G}}$ with respect to the restriction $P|_{\mathscr{G}}$. This $\mathscr{G}$-measurable random variable on $(\Omega, \mathscr{F}, P)$ is obtained by evaluating the Radon–Nikodym derivative of the absolutely continuous part of $Q|_{\mathscr{G}}$ relative to $P|_{\mathscr{G}}$, at the actual outcome ω.

Let $(\Omega, \mathscr{F})$ be equipped with two probability measures, a reference measure M that is νth order Markov with stationary transition kernel $M(dx_\nu|x_0^{\nu-1})$ and a stationary measure P that is the true distribution of the process $\{X_t\}$. The finite dimensional marginals of M are assumed to dominate the corresponding marginals of P, and $p(x_0, \ldots, x_{n-1})$ will designate the density of the restriction $P|_{\sigma(X_0^{n-1})}$ relative to the restriction $M|_{\sigma(X_0^{n-1})}$. Thus

$$p(X_0, \ldots, X_{n-1}) = \left.\frac{dP}{dM}\right|_{\sigma(X_0^{n-1})}. \tag{21}$$

Let P be extended to a stationary distribution on the two-sided sequence space $\mathscr{X}_{-\infty}^\infty$. We designate by R the probability measure on $\mathscr{X}_{-\infty}^\infty$ such that $X_{-\infty}^{-1}$ is distributed as under P and the transition kernel $R(dx_t|x_{-\infty}^{t-1})$ is a copy $M(dx_t|x_{t-\nu}^{t-1})$ of the transition kernel of M, for all $t \ge 0$. In particular $R|_{\sigma(X_{-\infty}^0)}$ is obtained by extension to $\sigma(X_{-\infty}^0)$ of the set function

$$R(B \times C) = \int_B M(C|x_{-\nu}^{-1})\, dP, \qquad B \in \sigma(X_{-\infty}^{-1}),\ C \in \sigma(X_0). \tag{22}$$

The finite dimensional marginals of P are dominated by the corresponding marginals of R. For finite $k \ge \nu$ let $p(x_0|x_{-1}, \ldots, x_{-k})$ denote the density of P relative to R after restriction to $\sigma(X_{-k}^0)$. It is well known [cf. Neveu (1972), Proposition III-2-7] that $\{p(X_0|X_{-k}^{-1}), \sigma(X_{-k}^0)\}_{\nu \le k < \infty}$ is an R-martingale, converging a.s. (R) to the density of the absolutely continuous part of $P|_{\sigma(X_{-\infty}^0)}$

relative to $R|_{\sigma(X_{-\infty}^0)}$:

$$p(X_0|X_{-k}^{-1}) = \left.\frac{dP}{dR}\right|_{\sigma(X_{-k}^0)} \to p(X_0|X_{-\infty}^{-1}) = \left.\frac{dP}{dR}\right|_{\sigma(X_{-\infty}^0)} \quad \text{a.s. } (R) \text{ as } k \to \infty. \tag{23}$$

Let $p(X_t|X_{t-k}^{t-1})$ denote the random variable obtained by shifting $p(X_0|X_{-k}^{-1})$ over t periods.

The *relative entropy rate* is defined as $I_M(P) = \lim_k \uparrow I_M^k(P)$, where

$$I_M^k(P) = E\{\log p(X_k|X_0^{k-1})\} = E\{\log p(X_0|X_{-k}^{-1})\} \quad \text{for } k \geq \nu. \tag{24}$$

If $I_M(P) < \infty$, then P is dominated by R on $\sigma(X_{-\infty}^0)$, with density $p(X_0|X_{-\infty}^{-1}) = \lim_k p(X_0|X_{-k}^{-1})$ a.s. (P), and $\{\log p(X_0|X_{-k}^{-1}), \sigma(X_{-k}^0)\}_{\nu \leq k < \infty}$ is a uniformly integrable P-submartingale. Consequently if the limit of expectations $I_M(P) = \lim_k \uparrow I_M^k(P)$ is finite, then it coincides with the expectation of the limit, i.e.,

$$I_M(P) = E\{\log p(X_0|X_{-\infty}^{-1})\} \quad \text{if } I_M(P) < \infty. \tag{25}$$

A proof of these facts is given in full in Moy (1961) and as Exercise IV-5-5 in Neveu (1970).

The AEP will follow from a lemma that may be of independent interest.

LEMMA 2 (Sandwich lemma). *Let* $\{\underline{Z}_n\}$, $\{Z_n\}$ *and* $\{\overline{Z}_n\}$ *be sequences of positive random variables.*

(a) *If* $\sup_n E\{\underline{Z}_n/Z_n\} < \infty$ *or more generally if* $E\{\underline{Z}_n/Z_n\}$ *has subexponential growth* (*i.e.*, $\limsup_n n^{-1}\log E\{\underline{Z}_n/Z_n\} \leq 0$), *then*

$$\liminf_n n^{-1}\log \underline{Z}_n \leq \liminf_n n^{-1}\log Z_n \quad a.s. \tag{26}$$

(b) *If* $\sup_n E\{Z_n/\overline{Z}_n\} < \infty$ *or more generally if* $E\{Z_n/\overline{Z}_n\}$ *has subexponential growth* (*i.e.*, $\limsup_n n^{-1}\log E\{Z_n/\overline{Z}_n\} \leq 0$), *then*

$$\limsup_n n^{-1}\log Z_n \leq \limsup_n n^{-1}\log \overline{Z}_n \quad a.s. \tag{27}$$

PROOF. Let $\overline{C}_n = E\{Z_n/\overline{Z}_n\}$ and suppose $\varepsilon > 0$. By Markov's inequality,

$$P\{n^{-1}\log(Z_n/\overline{Z}_n) > \varepsilon\} = P\{Z_n/\overline{Z}_n > \exp(n\varepsilon)\} \leq \overline{C}_n \exp(-n\varepsilon).$$

Since $\Sigma_n \overline{C}_n \exp(-n\varepsilon) < \infty$ for arbitrary $\varepsilon > 0$, the Borel–Cantelli lemma gives

$$\limsup_n n^{-1}\log(Z_n/\overline{Z}_n) \leq 0 \quad \text{a.s.}$$

One obtains the chain of asymptotic inequalities

$$\begin{aligned}\limsup_n n^{-1}\log Z_n &= \limsup_n \left[n^{-1}\log(Z_n/\overline{Z}_n) + n^{-1}\log \overline{Z}_n\right] \\ &\leq \limsup_n n^{-1}\log(Z_n/\overline{Z}_n) + \limsup_n n^{-1}\log \overline{Z}_n \\ &\leq \limsup_n n^{-1}\log \overline{Z}_n.\end{aligned}$$

This proves (b) and the proof of (a) is analogous. □

We obtain the AEP by applying the sandwich lemma to likelihood ratios, which have expectations bounded by 1.

THEOREM 2 (Generalized AEP for stationary ergodic P). *Suppose M is νth order Markov with stationary transition kernel $M(dx_\nu|x_0^{\nu-1})$, and the finite dimensional marginals of M dominate the corresponding marginals of a stationary measure P. If P is ergodic, then*

$$n^{-1}\log p(X_0,\dots,X_{n-1}) \to I_M(P) \quad a.s.\,(P). \tag{28}$$

PROOF. For finite $k \geq \nu$ let P^k designate the kth order Markov approximation of P, that is the stationary kth order Markov distribution on $\mathscr{X}_0^\infty$ having the same $(k+1)$st order marginals as P. If $\nu \leq k \leq n < \infty$, then P is dominated by M on $\sigma(X_0^{n-1})$ with likelihood ratio

$$\begin{aligned}\left.\frac{dP^k}{dM}\right|_{\sigma(X_0^{n-1})} &= p^k(X_0,\dots,X_{n-1}) \\ &= p(X_0,\dots,X_{k-1})\prod_{t=k}^{n-1} p(X_t|X_{t-1},\dots,X_{t-k}) \quad \text{a.s.}\,(P).\end{aligned} \tag{29}$$

By the chain rule for densities,

$$\left.\frac{dP^k}{dP}\right|_{\sigma(X_0^{n-1})} = \left.\frac{dP^k}{dM}\right|_{\sigma(X_0^{n-1})} \Bigg/ \left.\frac{dP}{dM}\right|_{\sigma(X_0^{n-1})} = \frac{p^k(X_0,\dots,X_{n-1})}{p(X_0,\dots,X_{n-1})} \quad \text{a.s.}\,(P). \tag{30}$$

This likelihood ratio has expectation no larger than 1. Using part (a) of Lemma 2 and the ergodic theorem we obtain the asymptotic lower bound

$$\begin{aligned}\liminf_n n^{-1}\log p(X_0,\dots,X_{n-1}) &\geq \lim_n n^{-1}\log p^k(X_0,\dots,X_{n-1}) \\ &= I_M^k(P) \quad \text{a.s.}\,(P).\end{aligned} \tag{31}$$

Now suppose $I_M(P) < \infty$, so that $p(X_0|X_{-\infty}^{-1})$ is a bona fide density and $I_M(P) = E\{\log p(X_0|X_{-\infty}^{-1})\}$. Let P^∞ denote the distribution on $\mathscr{X}_{-\infty}^\infty$ such that $X_{-\infty}^{\nu-1}$ is distributed as under P, and the conditional distribution $P^\infty(dx_t|X_{-\infty}^{t-1})$ is equal to $P(dx_t|X_0^{t-1})$ for all $t \geq \nu$. If $n \geq \nu$, then by the chain rule for conditional densities

$$\left.\frac{dP^\infty}{dP}\right|_{\sigma(X_{-\infty}^{n-1})} = \frac{p(X_0,\dots,X_{n-1})/p(X_0,\dots,X_{\nu-1})}{\prod_{t=\nu}^{n-1} p(X_t|X_{-\infty}^{t-1})} \quad \text{a.s.}\,(P). \tag{32}$$

This likelihood ratio also has expectation no larger than 1. Using part (b) of Lemma 2 and the ergodic theorem we obtain the asymptotic upper bound

$$\begin{aligned}\limsup_n n^{-1}\log p(X_0,\dots,X_{n-1}) &\leq E\{\log p(X_0|X_{-1},X_{-2},\cdots)\} \\ &= I_M(P) \quad \text{a.s.}\,(P).\end{aligned} \tag{33}$$

But $I_M^k(P) \nearrow I_M(P)$, so the AEP follows by chaining the asymptotic inequalities (31) and (33) if $I_M(P) < \infty$ and from (31) alone otherwise. □

Note that $I_M^k(P) = I_M(P^k)$ is the relative entropy rate, relative to M, of the kth order Markov approximation P^k of P. The difference $I_M(P) - I_M^k(P)$ can be interpreted as the mutual information $I(X_0; X_{-\infty}^{-k-1}|X_{-k}^{-1}) = I(X_0; X_{-\infty}^{-1}|X_{-k}^{-1})$, and also as the information divergence rate $I_{P^k}(P)$ of P relative to P^k. The AEP (28) with $M = P^k$ gives

$$\lim_n n^{-1}\log\left(\frac{p^k(X_0,\ldots,X_{n-1})}{p(X_0,\ldots,X_{n-1})}\right) = I_{P^k}(P) = I\left(X_0; X_{-\infty}^{-1}|X_{-k}^{-1}\right) \quad \text{a.s. } (P). \tag{34}$$

It is perhaps worthwhile to point out the relation between entropy rate and relative entropy rate. If μ is a (necessarily σ-finite) reference measure on a standard Borel space $\mathcal{X}$ and $\{X_t\}$ is an $\mathcal{X}$-valued stationary ergodic process with distribution P admitting conditional and joint densities $p(x_n|x_{n-1},\ldots,x_0)$ and $p(x_0,\ldots,x_{n-1})$ relative to μ (respectively, relative to the product of n copies of μ), then the entropy rate $H_\mu(P)$ is defined as $\lim_k \downarrow H_\mu^k(P)$, where $H_\mu^k(P) = E\{-\log p(X_k|X_0^{k-1})\}$. Although $H_\mu(P)$ is always nonnegative if μ is counting measure on a countable set $\mathcal{X}$, no such claim can be made in general, e.g., if μ is Lebesgue measure on the real line. Nevertheless,

$$\begin{aligned} -n^{-1}\log p(X_0,\ldots,X_{n-1}) &= -n^{-1}\sum_{t=0}^{n-1}\log p(X_t|X_{t-1},\ldots,X_0) \\ &\to H_\mu(P) \quad \text{a.s. } (P). \end{aligned} \tag{35}$$

If M is the product of copies of a distribution m that dominates μ on $\mathcal{X}$, then

$$H_\mu(P) = E\left\{\log\left(\frac{d\mu}{dm}(X_0)\right)\right\} - I_M(P). \tag{36}$$

In particular, if μ is counting measure on a finite set $\mathcal{X}$ with cardinality $\|\mathcal{X}\|$ and M is the product of copies of the normalized measure $m = \mu/\|\mathcal{X}\|$, then $I_M(P) = \log\|\mathcal{X}\| - H_\mu(P)$.

4. The AEP for stationary and asymptotically mean stationary P. Two generalizations of the AEP due to Barron (1985) will now be proved using the sandwich technique.

If P is stationary but not ergodic, then the σ-field of invariant events $\mathcal{I} = \{F \in \mathcal{F}: T^{-1}F = F\}$ is nontrivial. The relative entropy rate $I_M(P)$ is then the expectation of the invariant random variable $i_M = \lim_k \uparrow i_M^k$, where

$$i_M^k = E\left\{\log p\left(X_k|X_0^{k-1}\right)|\mathcal{I}\right\} = E\left\{\log p\left(X_0|X_{-k}^{-1}\right)|\mathcal{I}\right\} \quad \text{for } k \geq \nu. \tag{37}$$

To prove that i_M^k is nondecreasing it suffices to observe for $\nu \leq k \leq l \leq n$ that

$$E\left\{\frac{p^k(X_0,\ldots,X_{n-1})}{p^l(X_0,\ldots,X_{n-1})}\right\} \leq 1, \tag{38}$$

and hence, by Lemma 2 and the ergodic theorem,

$$\begin{aligned} i_M^k &= \lim_n n^{-1}\log p^k(X_0,\dots,X_{n-1}) \\ &\le i_M^l = \lim_n n^{-1}\log p^l(X_0,\dots,X_{n-1}) \quad \text{a.s. } (P). \end{aligned} \tag{39}$$

THEOREM 3 (Generalized AEP for stationary P). *If P is stationary but the other hypotheses of Theorem 2 hold, then*

$$n^{-1}\log p(X_0,\dots,X_{n-1}) \to i_M = \lim_k \uparrow i_M^k \quad a.s.\,(P). \tag{40}$$

PROOF. If $I_M(P) = E\{i_M\}$ is finite, then (40) follows by substitution of the invariant random variables $i_M^k = E\{\log p(X_0|X_{-k}^{-1})|\mathcal{I}\}$ and $i_M = E\{\log p(X_0|X_{-\infty}^{-1})|\mathcal{I}\}$ for the limiting expectations $I_M^k(P)$ and $I_M(P)$ in (31) and (33). If $I_M(P) = \infty$, then we define $\Omega_N = \{i_M < N\}$ for integer N and observe that $\{\Omega_N\}$ is an increasing sequence of invariant events such that the asymptotic lower bounds (31) (upgraded by writing i_M^k on the right) and hence (40) hold on the complement of $\bigcup_N \Omega_N$. But (40) holds on Ω_N for finite N such that $P(\Omega_N) > 0$, since it holds under the conditional measure $P(\cdot|\Omega_N) = P(\cdot \cap \Omega_N)/P(\Omega_N)$ and the constant $\log P(\Omega_N)$ may be added to or subtracted from both sides without making a difference. Thus (40) holds without restriction. □

A probability distribution P in $(\Omega, \mathcal{F})$ is called *asymptotically mean stationary* (a.m.s.) if the Cesàro averages $n^{-1}\sum_{t=0}^{n-1} P(T^{-t}F)$ converge for all Borel sets $F \in \mathcal{F}$. Setting the limit equal to $\overline{P}(F)$ then defines a stationary measure $\overline{P}$, which is called the stationary mean of P. Clearly $\overline{P}$ and P have the same restriction to the invariant σ-field $\mathcal{I}$, so that $E\{\cdot|\mathcal{I}\} = \overline{E}\{\cdot|\mathcal{I}\}$ if $\overline{E}\{\cdot\}$ denotes expectation with respect to $\overline{P}$. See Gray and Kieffer (1980) for further discussion of a.m.s. measures, and Section 34.2 in Loève (1978) for a proof that the following strong law of large numbers holds for nonnegative measurable $g(\omega)$:

$$n^{-1}\sum_{t=0}^{n-1} g(T^t\omega) \to E\{g|\mathcal{I}\} = \overline{E}\{g|\mathcal{I}\} \quad \text{a.s. } (P) \text{ and a.s. } (\overline{P}). \tag{41}$$

THEOREM 4 (Generalized AEP for asymptotically mean stationary P). *Suppose M is finite order Markov with stationary transition kernel and the finite dimensional marginals of M dominate the corresponding marginals of an asymptotically mean stationary measure P as well as those of its stationary mean $\overline{P}$. The AEP will hold for P with the same limiting rate as under $\overline{P}$ if $\limsup_n n^{-1}\log p(X_0,\dots,X_{n-1})$ is an invariant random variable. In particular if $\overline{P}$ is ergodic then the AEP for P asserts that*

$$n^{-1}\log p(X_0,\dots,X_{n-1}) \to I_M(\overline{P}) \quad a.s.\,(P). \tag{42}$$

PROOF. We consider the ergodic case. Let $\overline{p}^k(x_0,\dots,x_{n-1})$, $p(x_0,\dots,x_{n-1})$ and $\overline{p}(x_0,\dots,x_{n-1})$ denote the densities of $\overline{P}^k$, P and $\overline{P}$ relative to M after

restriction to $\sigma(X_0^{n-1})$. Then

$$(43)\qquad E\left\{\frac{\bar{p}^k(X_0,\ldots,X_{n-1})}{p(X_0,\ldots,X_{n-1})}\right\}\le 1 \quad\text{and}\quad \bar{E}\left\{\frac{p(X_0,\ldots,X_{n-1})}{\bar{p}(X_0,\ldots,X_{n-1})}\right\}\le 1.$$

Part (a) of Lemma 2 in conjunction with the ergodic theorem (41) proves that

$$(44)\qquad \begin{aligned}\liminf_n n^{-1}\log p(X_0,\ldots,X_{n-1}) &\ge \lim_n n^{-1}\log \bar{p}^k(X_0,\ldots,X_{n-1})\\ &= I_M(\bar{P}^k)\quad \text{a.s. } (P)\end{aligned}$$

and part (b) in conjunction with the AEP for the stationary mean $\bar{P}$ yields

$$(45)\qquad \begin{aligned}\limsup_n n^{-1}\log p(X_0,\ldots,X_{n-1}) &\le \lim_n n^{-1}\log \bar{p}(X_0,\ldots,X_{n-1})\\ &= I_M(\bar{P})\quad \text{a.s. } (\bar{P}).\end{aligned}$$

If $\limsup_n n^{-1}\log p(X_0,\ldots,X_{n-1})$ is an invariant random variable, then (45) holds not only a.s. $(\bar{P})$ but also a.s. (P) and the AEP (42) follows. □

Notice that $\limsup_n n^{-1}\log p(X_0,\ldots,X_{n-1})$ is invariant if a sequence $\{k_n\}$ exists such that $k_n\to\infty$, $k_n/n\to 0$, and $n^{-1}\log p(X_0,\ldots,X_{k_n})\to 0$ a.s. (P). Barron (1985) put forward this condition in his Theorem 3, and reduced it to existence of $\{m_k\}$ such that the mutual information $I(X_0^{k-1};\, X_k^{\infty}|X_k^{k+m_k-1})$ is finite for all $k\ge 1$.

Algoet and Cover (1988) provides a gambling interpretation of the AEP and further motivation of the sandwich argument.

Acknowledgments. We thank A. R. Barron, R. M. Gray and J. C. Kieffer for helpful comments. We are especially grateful to the referee for suggesting Lemma 2 and for bringing Orey's (1985) paper to our attention.

REFERENCES

ALGOET, P. H. and COVER, T. M. (1988). Asymptotic optimality and asymptotic equipartition properties of log-optimum investment. *Ann. Probab.* **166** 876–898.

BARRON, A. R. (1985). The strong ergodic theorem for densities: Generalized Shannon–McMillan–Breiman theorem. *Ann. Probab.* **13** 1292–1303.

BREIMAN, L. (1957/1960). The individual ergodic theorem of information theory. *Ann. Math. Statist.* **28** 809–811. Correction **31** 809–810.

CHERNOFF, H. (1956). Large sample theory—parametric case. *Ann. Math. Statist.* **27** 1–22.

CHUNG, K. L. (1961). A note on the ergodic theorem of information theory. *Ann. Math. Statist.* **32** 612–614.

CHUNG, K. L. (1962). The ergodic theorem of information theory. In *Recent Developments in Information and Decision Processes* 141–148. Macmillan, New York.

GRAY, R. M. and KIEFFER, J. C. (1980). Asymptotically mean stationary measures. *Ann. Probab.* **8** 962–973.

KIEFFER, J. C. (1973/1976). A counterexample to Perez's generalization of the Shannon–McMillan theorem. *Ann. Probab.* **1** 362–364. Correction **4** 153–154.

KIEFFER, J. C. (1974). A simple proof of the Moy–Perez generalization of the Shannon–McMillan theorem. *Pacific J. Math.* **51** 203–206.

LOÈVE, M. (1978). *Probability Theory* **2**, 4th ed. Springer, New York.
MCMILLAN, B. (1953). The basic theorems of information theory. *Ann. Math. Statist.* **24** 196–219.
MOY, S. C. (1960). Asympotitic properties of derivatives of stationary measures. *Pacific J. Math.* **10** 1371–1383.
MOY, S. C. (1961). Generalizations of Shannon–McMillan theorem. *Pacific J. Math.* **11** 705–714.
NEVEU, J. (1970). *Calcul des Probabilités*. Masson, Paris.
NEVEU, J. (1972). *Martingales à Temps Discret*. Masson, Paris.
OREY, S. (1985). On the Shannon–Perez–Moy theorem. *Contemp. Math.* **41** 319–327.
PEREZ, A. (1964). Extensions of Shannon–McMillan's limit theorem to more general stochastic processes. In *Trans. Third Prague Conf. Inform. Theory, Statist. Decision Functions and Random Processes* 545–574. Czechoslovak. Acad. Sci., Prague.
PEREZ, A. (1974). Generalization of Chernoff's result on the asymptotic discernibility of two random processes. *Colloq. Math. Soc. János Bolyai* **9** 619–632.
SHANNON, C. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27** 379–423, 623–656.

COLLEGE OF ENGINEERING
BOSTON UNIVERSITY
110 CUMMINGTON STREET
BOSTON, MASSACHUSETTS 02215

DEPARTMENTS OF STATISTICS AND
ELECTRICAL ENGINEERING
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305