The Strong Ergodic Theorem for Densities: Generalized Shannon-McMillan-Breiman Theorem

Andrew R. Barron

The Annals of Probability, Vol. 13, No. 4 (Nov., 1985), 1292-1303.

Stable URL:
http://links.jstor.org/sici?sici=0091-1798%28198511%2913%3A4%3C1292%3ATSETFD%3E2.0.CO%3B2-J

The Annals of Probability is currently published by Institute of Mathematical Statistics.

The Annals of Probability
1985, Vol. 13, No. 4, 1292–1303

THE STRONG ERGODIC THEOREM FOR DENSITIES: GENERALIZED SHANNON–MCMILLAN–BREIMAN THEOREM[1]

BY ANDREW R. BARRON

Stanford University

Let $\{X_1, X_2, \dots\}$ be a stationary process with probability densities $f(X_1, X_2, \dots, X_n)$ with respect to Lebesgue measure or with respect to a Markov measure with a stationary transition measure. It is shown that the sequence of relative entropy densities $(1/n)\log f(X_1, X_2, \dots, X_n)$ converges almost surely. This long-conjectured result extends the L^1 convergence obtained by Moy, Perez, and Kieffer and generalizes the Shannon–McMillan–Breiman theorem to nondiscrete processes. The heart of the proof is a new martingale inequality which shows that logarithms of densities are L^1 dominated.

1. Introduction. Let $(\Omega, \mathcal{B}, P)$ be a probability space and let $\{X_1, X_2, \dots\}$ be a stochastic process with each X_n taking values in a standard Borel space. Suppose that the joint distribution P_n for $(X_1, X_2, \dots, X_n)$ has a probability density function $f(X_1, X_2, \dots, X_n)$ with respect to a sigma-finite measure M_n. Assume that the sequence of dominating measures M_n is Markov of order $m \geq 0$ with a stationary transition measure. Familiar cases for M_n are Lebesgue measure, counting measure, or a Markov probability measure serving as an alternative in a hypothesis test. Let $f(X_{n+1}|X_1, \dots, X_n)$ denote the conditional density given by the ratio $[f(X_1, \dots, X_{n+1})]/[f(X_1, \dots, X_n)]$ for $n \geq 1$, and by $f(X_1)$ for $n = 0$. Let E denote expectation with respect to P and let log be the natural logarithm. If $\{X_n\}$ is stationary, then the relative entropy $D_n = E \log f(X_{n+1}|X_1, \dots, X_n)$ is nondecreasing for $n \geq m$. (Indeed, if $D_n > -\infty$, the difference $D_{n+1} - D_n$ is a mutual information which is nonnegative by the concavity of the log.) The limit $D = \lim D_n$ is called the relative entropy rate. We are interested in the asymptotic behavior of the density $f(X_1, X_2, \dots, X_n)$. In particular, what is the exponential rate of growth? Our main result is the following.

THEOREM 1. *If $\{X_n\}$ is a P-stationary ergodic process with densities $f(X_1, X_2, \dots, X_n) = dP_n/dM_n$, and if $D_n > -\infty$ for some $n \geq m$, the sequence of relative entropy densities $(1/n)\log f(X_1, X_2, \dots, X_n)$ converges almost surely to the relative entropy rate* i.e.,

$$\lim \frac{1}{n} \log f(X_1, X_2, \dots, X_n) = D \quad P\text{-a.s.}$$

Received September 1983; revised December 1984.

[1]This work was partially supported by NSF Contract ECS 82-11568.

AMS 1980 *subject classifications.* Primary 28D05, 94A17; secondary 62B10, 60F15, 60G10, 60G42, 28D20.

Key words and phrases. Shannon–McMillan–Breiman theorem, Moy–Perez theorem, asymptotic equipartition property, martingale inequalities, entropy, information, ergodic theorems, asymptotically mean stationary.

The condition $D_n > -\infty$ is automatically satisfied if M_n is a finite measure (since $f \log f \geq -e^{-1}$). Indeed, if M_n is a probability measure then D_n is nonnegative.

Theorem 1 extends the L^1 convergence obtained by Moy (1961), Perez (1964), and Kieffer (1974). Moreover, it generalizes the Shannon–McMillan–Breiman theorem which asserts almost sure and L^1 convergence for discrete processes with counting measure for M_n [see Shannon (1948), McMillan (1953), Breiman (1957, 1960), Carleson (1958), Chung (1961, 1962), and Parthasarathy (1964)]. For L^p convergence with $p > 1$, see Ionescu Tulcea (1960).

The proof of Theorem 1 is given in Section 3 and uses Breiman's (1957) generalized ergodic theorem. The key to the proof is a new inequality for logarithms of supermartingales, derived in Section 2. In Sections 4 and 5 we extend our result to nonergodic and asymptotically mean stationary processes. Analogous convergence results for mutual information densities are obtained in Section 6.

Kieffer (1973, 1976) has provided counterexamples showing that without the Markov property for the dominating measures M_n, the sequence of relative entropy densities need not converge.

The following remarks indicate some applications of results such as Theorem 1. Let $\{X_n\}$ be a stationary ergodic process with P_n and M_n as in Theorem 1 and suppose the relative entropy rate D is finite. From convergence in probability of $(1/n)\log f(X_1, X_2, \ldots, X_n)$ to D, it follows that the least asymptotic M_n measure of P_n-probable sets A_n, is given by $M_n(A_n) \doteq e^{-nD}$ [specifically, $\lim P_n(A_n) = 1$ implies $\liminf (1/n)\log M_n(A_n) \geq -D$] and this rate is attained when A_n is the set of typical sequences with density near e^{nD}. (A sequence $(X_1, X_2, \ldots, X_n)$ is said to be typical if the density $f(X_1, X_2, \ldots, X_n)$ is between $e^{n(D-\epsilon)}$ and $e^{n(D+\epsilon)}$, where ϵ may tend slowly to zero as $n \to \infty$.) For a hypothesis test between distributions P_n and M_n, this property is a generalized Stein's lemma for the best exponent in the probability of error [see Chernoff (1956), p. 17 for the stationary independent case]. In the information theory context, McMillan (1953) called this property the AEP (asymptotic equipartition property). For processes with densities with respect to Lebesgue measure, the AEP states that the distribution P_n is asymptotically uniform on the set A_n of typical sequences and that this set has the least asymptotic volume e^{-nD} among sets of high probability. Theorem 1 ensures that random sequences $(X_1, X_2, \ldots, X_n)$ are indeed typical, for all large n, with probability one (not only in probability).

Our convergence result has potential applications to the theory of statistical inference. Let $\{M^\theta : \theta \in \Theta\}$ be a parametric family of Markov probability measures with stationary transition probabilities, and let $l_n(\theta) = dM_n^\theta/d\mu_n$ be the likelihood functions with respect to sigma-finite measures μ_n. Assume that the true but unknown distribution P with densities $p_n = dP_n/d\mu_n$ is stationary and ergodic but not necessarily Markov. Events that order the likelihoods $l_n(\theta)$ (for instance $\{l_n(\theta_1) > l_n(\theta_2)\}$) are the same as events that order the unknown relative entropy densities $(1/n)\log p_n/l_n(\theta)$. These relative entropy densities converge P-a.s. to the corresponding relative entropy rate D^θ. In particular, if the parameter space Θ is a finite set, then with probability one, for all large n, the estimate θ_{ML} which maximizes the likelihood is equal to that θ for which the

realtive entropy rate D^θ is minimum. Similar arguments have been used in the stationary independent case (where convergence of the relative entropy densities reduces to the ordinary law of large numbers) to yield the classic proofs of the following results: the consistency of maximum likelihood parameter estimates [see Wald (1949)], the consistency of Bayes estimates [see Schwartz (1965)], and the exact slope of the likelihood ratio statistic [see Bahadur (1971)].

2. Logarithms of densities are L^1 dominated. In this section we establish a supermartingale inequality and apply it to sequences of densities.

Let $\{Z_n\}$ be a sequence of nonnegative real-valued random variables adapted to an increasing sequence of sigma fields. Let $\log^+$ and $\log^-$ denote the positive and negative parts of the natural logarithm.

LEMMA 1. *If $\{Z_n\}$ is a nonnegative supermartingale, then*

$$E \sup_n \log^- Z_n \le (1 + e) + e \sup_n E \log^- Z_n. \tag{2.1}$$

PROOF. An inequality due to Doob (1953, p. 317) asserts that if $\{Y_n\}$ is a nonnegative submartingale, then for any $r > 1$

$$E \sup_n Y_n^r \le \left(\frac{r}{r-1}\right)^r \sup_n E Y_n^r. \tag{2.2}$$

Fix $r > 1$ and define $Y_n = \phi(Z_n) = \max\{1, (\log^- Z_n)^{1/r}\}$. The function ϕ is nonincreasing and convex, so $\{Y_n\}$ is a nonnegative submartingale. From Doob's inequality we obtain

$$E \sup_n \log^- Z_n \le 1 + \left(\frac{r}{r-1}\right)^r \left(1 + \sup_n E \log^- Z_n\right). \tag{2.3}$$

Taking the limit as $r \to \infty$ completes the proof of Lemma 1. □

The next lemma shows that logarithms of densities are L^1 dominated. Let Q and R be probability measures on a measurable space and let Q_n and R_n be the restrictions to an increasing sequence of sigma fields $\mathcal{F}_n$. Expectations are taken with respect to Q.

LEMMA 2. *If the densities $\rho_n = dQ_n/dR_n$ exist, then $\rho_\infty = \lim \rho_n$ exists Q-a.s., $E \log \rho_n$ increases to $E \log \rho_\infty$, $E \sup_n \log^- \rho_n \le 1$, and*

$$E \sup_n |\log \rho_n| \le eE \log \rho_\infty + (e + 3). \tag{2.4}$$

PROOF. The density sequence $\{\rho_n\}$ is an R martingale. Thus by concavity $\{I_{\{\rho_n > 0\}}\}$ is an R supermartingale, or equivalently, $\{1/\rho_n\}$ is a Q supermartingale with expectations $E_Q(1/\rho_n) = E_R I_{\{\rho_n > 0\}}$ bounded by one. Hence ρ_n converges Q-a.s. to a (possibly infinite) limit ρ_∞. By Lemma 1 we have

$$E \sup_n \log^- 1/\rho_n \leq e \sup_n E \log^- 1/\rho_n + (e+1). \tag{2.5}$$

On the other hand,

$$E \sup_n \log^+ 1/\rho_n \leq 1. \tag{2.6}$$

Inequality (2.6) is due to Ionescu Tulcea (1960) and the proof is as follows. The expected supremum equals $\int_0^\infty Q(A_t)\,dt$ where A_t is the event $\{\sup_n(\log 1/\rho_n) > t\}$. Write A_t as the disjoint union of events $A_{t,n} = \{\log 1/\rho_n > t, \max_{k<n} \log 1/\rho_k \leq t\}$. Then $Q(A_{t,n}) = \int_{A_{t,n}} \rho_n\,dR$. But within $A_{t,n}$, the density satisfies $\rho_n < e^{-t}$, so that $Q(A_{t,n}) \leq e^{-t} R(A_{t,n})$. Summing over n yields $Q(A_t) \leq e^{-t}$. Integrating yields (2.6).

The relative entropy satisfies $E_Q \log \rho_n = E_R \rho_n \log \rho_n$ (if we define $0 \log 0 = 0$). From $\rho_n \log \rho_n \geq -e^{-1}$ we obtain $E \log^+ \rho_n \leq E \log \rho_n + e^{-1}$. Therefore inequalities (2.5) and (2.6) yield

$$E \sup_n |\log \rho_n| \leq e \sup_n E \log \rho_n + (e+3). \tag{2.7}$$

Now by concavity $\{\log \rho_n\}$ is a Q submartingale and hence has nondecreasing expectation $E \log \rho_n$. If $E \log \rho_n$ is bounded, (2.7) shows that $\log \rho_n$ is L^1 dominated, so by the dominated convergence theorem

$$\lim_n E \log \rho_n = \sup_n E \log \rho_n = E \log \rho_\infty. \tag{2.8}$$

From (2.6) $E \sup_n \log^- \rho_n \leq 1$, so even if $\sup_n E \log \rho_n = \infty$, (2.8) holds by Fatou's lemma. This completes the proof of Lemma 2. □

Note that if the relative entropy sequence $E \log \rho_n$ is bounded above, then the densities ρ_n are uniformly R-integrable (because $E_R \rho_n I_{\{\rho_n > \tau\}} \leq (E_R \rho_n \log^+ \rho_n)/\log \tau \leq (E_Q \log \rho_n + e^{-1})/\log \tau$ which tends to zero uniformly in n as $\tau \to \infty$), in which case Q is absolutely continuous with respect to R on the limit sigma-field $\mathcal{F}_\infty$ with density given by ρ_∞. (Because if A is any set in $\mathcal{F}_n$ for for any n, then $Q(A) = \lim Q_n(A) = \lim \int_A \rho_n\,dR = \int_A \lim \rho_n\,dR = \int_A \rho_\infty\,dR$, but the sequence $\mathcal{F}_n$ generates $\mathcal{F}_\infty$, so $Q(A) = \int_A \rho_\infty\,dR$ for any event A in $\mathcal{F}_\infty$.)

Now we recall some basic properties of conditional distributions and mutual information densities of random variable.

Let $(\Omega, \mathcal{B}, P)$ be a probability space. We assume throughout that the random variables U, V, W_1, W_2, etc. take values in standard Borel spaces. A measurable space is standard if it is (sigma isomorphic to) a Borel subspace of a complete separable metric space [see Parthasarathy (1967)]. The standard space assumption guarantees that conditional probability distributions $P_{U|V}$ exist. Specifically, $P_{U|V}$ is a probability measure for almost every V, it is a measurable function of V for each U-measurable event, and it integrates to give the unconditional probabil-

ity $P_{UV}(A) = \int P_{U|V}(A_V)\,dP_V$ for each event A in the product sigma field where $A_V = \{U: (U, V) \in A\}$ is the section of A at V.

Let U, V, and W be random variables. The conditional mutual information density is defined as $i(U; W|V) = \log \rho(U; W|V)$ where ρ is the conditional density of $P_{UW|V}$ with respect to $P_{U|V} \times P_{W|V}$ if this density exists, otherwise set $i = \infty$. The conditional density $\rho(U; W|V)$ is almost surely the same as the unconditional density of P_{UVW} with respect to $P_{U|V}P_{VW}$ [see Pinsker (1964), Section 3.1]. Therefore, an equivalent exprssion for the information density is

$$i(U; V|W) = \log \frac{dP_{UVW}}{d(P_{U|V}P_{VW})}. \tag{2.9}$$

Here $P_{U|V}P_{VW}$ is the set function (on the product sigma field) defined by $(P_{U|V}P_{VW})(A) = \int P_{U|V}(A_{VW})\,dP_{VW}$ where A_{VW} is the section of A at VW. (Note that $P_{U|V}P_{VW}$ has transition probability $P_{U|V}$ instead of $P_{U|VW}$ from VW to U.) Because $P_{U|V}$ is a probability measure for almost every V, it follows that $P_{U|V}P_{VW}$ is also a probability measure [as noted in Feinstein's translation of Pinsker (1964), p. 55].

The conditional mutual information is defined by $I(U; W|V) = Ei(U; W|V)$. If V is degenerate, this reduces to ordinary mutual information $I(U; W) = Ei(U; W) = E \log dP_{UW}/d(P_U \times P_W)$. Well known properties of conditional mutual information include positivity $I(U; W|V) \geq 0$ (with equality if and only if U and W are conditionally independent given V) and the chain rule $I(U; W_1, W_2|V) = I(U; W_1|V) + I(U; W_2|W_1, V)$. The chain rule also holds pointwise for the information densities.

For convenience, we use $\mathbf{W}_{k,n}$ to denote segments of a sequence of random variables $\{W_n\}$. Specifically let $\mathbf{W}_{k,n} = (W_k, W_{k+1}, \ldots, W_n)$ for $k \leq n$. Similarly define $\mathbf{W}_{k,\infty} = (W_k, W_{k+1}, \ldots)$ and $\mathbf{W}_{-\infty,n} = (\ldots, W_{n-1}, W_n)$. For $k > n$, regard $\mathbf{W}_{k,n}$ as degenerate.

The following lemma shows that sequences of information densities are L^1 dominated.

LEMMA 3. *Let $U, V, W_1, W_2, \ldots$ be random variables. The sequence of conditional mutual information densities $i(U; \mathbf{W}_{1,n}|V)$ converges almost surely, $I(U; \mathbf{W}_{1,n}|V)$ increases to $I(U; \mathbf{W}_{1,\infty}|V)$, $E \sup_n (i(U; \mathbf{W}_{1,n}|V))^- \leq 1$, and*

$$E \sup_n |i(U; \mathbf{W}_{1,n}|V)| \leq eI(U; \mathbf{W}_{1,\infty}|V) + (e + 3). \tag{2.10}$$

PROOF. This result follows from Lemma 2 with $Q = P_{UV\mathbf{W}_{1,\infty}}$, $R = P_{U|V}P_{V\mathbf{W}_{1,\infty}}$, and $\mathcal{F}_n$ the sigma field generated by $U, V, \mathbf{W}_{1,n}$. □

3. Proof of Theorem 1. Given the one-sided stationary process $\{X_1, X_2, \ldots\}$ with distributions P_n for $\mathbf{X}_{1,n}$, we extend it to a two-sided stationary process $\{\ldots, X_{-1}, X_0, X_1, \ldots\}$ in the usual way. In particular, we let $(\Omega, \mathcal{B}, P)$ be the probability space consisting of two-sided infinite sequences with the product sigma field and let P be given by the extension of the distributions $P_{\mathbf{X}_{i,j}} = P_{j-i+1}$ with $i \leq j$. Let the transformation T: $\Omega \to \Omega$ be the left shift.

By stationarity the conditional density $\mathrm{f}(X_{n+1}|\mathbf{X}_{1,n})$ (regarded as a function of the infinite sequence ω) equals the composition of $\mathrm{f}(X_0|\mathbf{X}_{-n,-1})$ with $n+1$ shifts. Therefore, the relative entropy density has the expansion

$$\frac{1}{n}\log \mathrm{f}(\mathbf{X}_{1,n}) = \frac{1}{n}\sum_{j=0}^{n-1} \log \mathrm{f}(X_0|\mathbf{X}_{-j,-1}) \circ T^{j+1}. \tag{3.1}$$

The almost sure (and L^1) convergence of (3.1) follows by Breiman's (1957) generalized ergodic theorem whenever the sequence $\log \mathrm{f}(X_0|\mathbf{X}_{-n,-1})$ is a.s. convergent and L^1 dominated.

Let $m \geq 0$ be the Markov order of M_n. By assumption there is a $k \geq m$ such that $D_k > -\infty$. By stationarity $D_n = E \log \mathrm{f}(X_{n+1}|\mathbf{X}_{1,n}) = E \log \mathrm{f}(X_0|\mathbf{X}_{-n,-1})$. From the Markov property for M_n and the chain rule for densities we have for $n \geq k$

$$\log \mathrm{f}(X_0|\mathbf{X}_{-n,-1}) = \log \mathrm{f}(X_0|\mathbf{X}_{-k,-1}) + i(X_0; \mathbf{X}_{-n,-k-1}|\mathbf{X}_{-k,-1}). \tag{3.2}$$

Hence by Lemma 3 the sequence $\mathrm{f}(X_0|\mathbf{X}_{-n,-1})$ converges a.s. [let $\mathrm{f}(X_0|\mathbf{X}_{-\infty,-1})$ denote the limit] and the relative entropy D_n increases to the relative entropy rate given by

$$D = D_k + I(X_0; \mathbf{X}_{-\infty,-k-1}|\mathbf{X}_{-k,-1}) = E \log \mathrm{f}(X_0|\mathbf{X}_{-\infty,-1}). \tag{3.3}$$

If this relative entropy rate is finite, then by (3.2) and Lemma 3 the sequence $\log \mathrm{f}(X_0|\mathbf{X}_{-n,-1})$ is L^1 dominated and hence by Breiman's theorem

$$\lim_n \frac{1}{n}\log \mathrm{f}(\mathbf{X}_{1,n}) = D \quad P\text{-a.s.} \tag{3.4}$$

To handle the infinite entropy rate case, a lim inf version of Breiman's theorem is used. In general, if $(\Omega, \mathscr{B}, P, T)$ is stationary and ergodic and if a sequence of random variables g_n is dominated from below ($E \sup_n g_n^- < \infty$), then $\liminf (1/n)\sum_{j=0}^{n-1} g_j \circ T^j \geq E \liminf g_n$ P-a.s. In particular, $g_n = \log \mathrm{f}(X_0|\mathbf{X}_{-n,-1})$ is dominated from below. Hence $\liminf (1/n)\log \mathrm{f}(\mathbf{X}_{1,n}) \geq D$ P-a.s. Therefore, even if the relative entropy rate is infinite, the sequence of relative entropy densities converges as in (3.4). This completes the proof of Theorem 1.

Even if $D_n = -\infty$ for all n, convergence as in (3.4) still holds if we assume $I(X_0; \mathbf{X}_{-\infty,-n-1}|\mathbf{X}_{-n,-1})$ is finite for some n [so that $\log \mathrm{f}(X_0|\mathbf{X}_{-n,-1})$ is dominated above]. Therefore, it is enough that either $D_n > -\infty$ or $I(X_0; \mathbf{X}_{-\infty,-n-1}|\mathbf{X}_{-n,-1}) < \infty$ for some n.

We remark that finite relative entropy rate implies a form of asymptotic conditional independence. Indeed, by (3.3) if the relative entropy D_n tends to a finite limit, then the conditional mutual information, $I(X_0; \mathbf{X}_{-\infty,-n-1}|\mathbf{X}_{-n,-1}) = D - D_n$, decreases to zero as $n \to \infty$, which means that the present X_0 is asymptotically conditionally independent of the far past $\mathbf{X}_{-\infty,-n-1}$ given the recent past $\mathbf{X}_{-n,-1}$.

Can Markov processes closely approximate non-Markov stationary processes? Let the Markov distribution M be constructed from P by using $P_{\mathbf{X}_{1,m}}$ as the initial distribution and $P_{X_{m+1}|\mathbf{X}_{1,m}}$ as the stationary transition probability. Then the relative entropy rate is simply the conditional mutual information

$I(X_0; \mathbf{X}_{-\infty,-m-1}|\mathbf{X}_{-m,-1})$, which is near zero for sufficiently large Markov order m (provided it is finite for some m). Thus the distribution P is approximated by the Markov distribution M in the relative entropy rate sense.

4. Extension to nonergodic processes. In this section we investigate the convergence of relative entropy densities $(1/n)\log f(\mathbf{X}_{1,n})$ for stationary but not necessarily ergodic processes. We find that the sequence of relative entropy densities converges almost surely to a shift invariant random variable. However, the proof requires the additional assumption that the relative entropy rate is finite.

Breiman's generalized ergodic theorem is readily extended to nonergodic processes. We find that if $(\Omega, \mathcal{B}, P, T)$ is stationary and if a sequence of random variables g_n is L^1 dominated and a.s. convergent with limit random variable g, then the sequence $(1/n)\sum_{j=0}^{n-1} g_j \circ T^j$ converges almost surely and in L^1 to the conditional expectation $E^{\mathcal{I}}g$. Here $\mathcal{I}$ is the sigma field of invariant events (events A such that $T^{-1}A = A$).

Set $g_n = \log f(X_0|\mathbf{X}_{-n,-1})$. Section 3 established that if the sequence $D_n = E\log f(X_0|\mathbf{X}_{-n,-1})$ is bounded [such that the relative entropy rate $D = E\log f(X_0|\mathbf{X}_{-\infty,-1})$ is finite], then the sequence $\log f(X_0|\mathbf{X}_{-n,-1})$ is L^1 dominated and almost surely convergent with limit given by $\log f(X_0|\mathbf{X}_{-\infty,-1})$. Applying the generalized ergodic theorem, we conclude that the sequence $(1/n)\log f(\mathbf{X}_{1,n})$ converges almost surely and in L^1 to $E^{\mathcal{I}}\log f(X_0|\mathbf{X}_{-\infty,-1})$. Thus we have established the following result.

THEOREM 2. *If $\{X_n\}$ is a P-stationary process with densities $f(\mathbf{X}_{1,n}) = dP_n/dM_n$, where M_n is a Markov measure with stationary transitions, and if the relative entropy sequence D_n is bounded, then the sequence of relative entropy densities $(1/n)\log f(\mathbf{X}_{1,n})$ converges P-almost surely and in $L^1(P)$ to the shift invariant random variable $E^{\mathcal{I}}\log f(X_0|\mathbf{X}_{-\infty,-1})$.*

Why do we require that the relative entropy rate be finite? If $E\log f(X_0|\mathbf{X}_{-\infty,-1}) = \infty$, then the sequence $\log f(X_0|\mathbf{X}_{-n,-1})$ is dominated from below but not from above. For the ergodic case the lim inf argument in Section 3 established that the limit is infinite. Applying the same technique to the nonergodic case yields $\liminf(1/n)\log f(\mathbf{X}_{1,n}) \geq E^{\mathcal{I}}\log f(X_0|\mathbf{X}_{-\infty,-1})$ a.s. The limit is not resolved in the case that the expectation $E\log f(X_0|\mathbf{X}_{\infty,1})$ is infinite but the conditional expectation is finite.

5. Asymptotic stationarity. In this section we generalize our density convergence result to asymptotic mean stationarity.

A probability space $(\omega, \mathcal{B}, P)$ together with a measurable transformation T: $\Omega \to \Omega$ is called asymptotically mean stationary if $(1/n)\sum_{j=0}^{n-1} P(T^{-j}A)$ is convergent for all $A \in \mathcal{B}$ [see Gray and Kieffer (1980a)]. Suppose that $(\Omega, \mathcal{B})$ is the one-sided sequence space for random variables $(X_1, X_2, \ldots)$ and that T is the left shift. Let $\mathcal{F}_\infty$ be the tail sigma field: that is, the intersection of the sigma fields generated by $(X_n, X_{n+1}, \ldots)$. A probability measure $\bar{P}$ is said to asymptoti-

cally dominate P if $A \in \mathscr{B}$ and $\bar{P}(A) = 0$ implies $\lim P(T^{-n}A) = 0$. Gray and Kieffer (1980a) demonstrated several useful properties of asymptotic stationarity and asymptotic dominance. In particular, if $\bar{P}$ is stationary then asymptotic dominance is equivalent to absolute continuity of the measures restricted to the tail sigma field [if $A \in \mathscr{F}_\infty$ and $\bar{P}(A) = 0$ then $P(A) = 0$]. If P is asymptotically dominated by a stationary measure, then P is asymptotically mean stationary. Conversely, if P is asymptotically mean stationary, then $\bar{P} = \lim(1/n)\sum_{j=0}^{n-1} PT^{-j}$ is a stationary measure that asymptotically dominates P. Furthermore, asymptotic mean stationarity is necessary and sufficient for an ergodic theorem to hold.

The following theorem generalizes the result of Gray and Kieffer (1980a) who treated the discrete case with counting measure for M_n. The generalization is made possible by our Lemma 3.

THEOREM 3. *Let P be an asymptotically mean stationary distribution for a stochastic process $\{X_1, X_2, \dots\}$. Suppose that for each $k \geq 1$ there exists an $m = m(k)$ such that $I_P(\mathbf{X}_{1,k}; \mathbf{X}_{k+m+1,\infty}|\mathbf{X}_{k+1,k+m})$ is finite. Let $\bar{P}$ be any stationary distribution which asymptotically dominates P. If the distributions P_n and $\bar{P}_n$ for $\mathbf{X}_{1,n}$ have densities $\mathrm{f}(\mathbf{X}_{1,n})$ and $\bar{\mathrm{f}}(\mathbf{X}_{1,n})$ with respect to M_n (a sigma finite Markov measure with stationary transitions), and if for some shift-invariant random variable D*

$$\frac{1}{n}\log \bar{\mathrm{f}}(\mathbf{X}_{1,n}) \to D \quad \bar{P}\text{-a.s.}, \tag{5.1}$$

then also

$$\frac{1}{n}\log \mathrm{f}(\mathbf{X}_{1,n}) \to D \quad P\text{-a.s.} \tag{5.2}$$

PROOF. First we show that there exists a subsequence $k(n) \to \infty$ sufficiently slowly that

$$\frac{1}{n}\log \mathrm{f}(\mathbf{X}_{1,n}) - \frac{1}{n}\log \mathrm{f}(\mathbf{X}_{k(n),n}) \to 0 \quad P\text{-a.s.} \tag{5.3}$$

For fixed k, $I(\mathbf{X}_{1,k}; \mathbf{X}_{k+m+1,\infty}|\mathbf{X}_{k+1,k+m})$ is decreasing in m, so we may assume that $m(k)$ exceeds the Markov order of M_n. Then for $S_k = \sup_n|\log \mathrm{f}(\mathbf{X}_{1,k}|\mathbf{X}_{k+1,n})|$ the chain rule for densities yields $S_k \leq |\log \mathrm{f}(\mathbf{X}_{1,k}|\mathbf{X}_{k+1,k+m})| + \sup_n|i(\mathbf{X}_{1,k}; \mathbf{X}_{k+m+1,n}|\mathbf{X}_{k+1,k+m})|$, where the supremum is over all $n > k + m(k)$. By Lemma 3, the expected supremum is finite and hence S_k is finite P-a.s. Therefore we can choose $k(n) \to \infty$ slowly enough that $S_{k(n)}/n \to 0$ P-a.s. [Specifically, choose increasing $c(k)$ large enough that $P\{S_k/c(k) > 2^{-k}\} < 2^{-k}$ and let $k(n) = k$ for $c(k) \leq n < c(k+1)$. From the Borel–Cantelli lemma $\lim_k S_k/c(k) = 0$ and hence $\lim S_{k(n)}/n = 0$ P-a.s.] Therefore, $\limsup(1/n)|\log \mathrm{f}(\mathbf{X}_{1,n})/\mathrm{f}(\mathbf{X}_{k(n),n})| \leq \limsup S_{k(n)}/n = 0$ P-a.s. which verifies (5.3).

Similarly, since information densities are dominated below, $k(n)$ can be chosen such that $\liminf(1/n)\log \bar{\mathrm{f}}(\mathbf{X}_{1,n})/\bar{\mathrm{f}}(\mathbf{X}_{k(n),n}) \geq 0$ $\bar{P}$-a.s. With (5.1) this

yields

$$\limsup \frac{1}{n} \log \bar{f}(\mathbf{X}_{k(n),n}) \le D \quad \bar{P}\text{-a.s.} \tag{5.4}$$

From Markov's inequality $\bar{P}\{\bar{f}(\mathbf{X}_{k(n),n}) \le f(\mathbf{X}_{k(n),n})e^{-n\epsilon}\} < e^{-n\epsilon}$ and hence by the Borel–Cantelli lemma $\liminf(1/n)\log \bar{f}(\mathbf{X}_{k(n),n})/f(\mathbf{X}_{k(n),n}) \ge 0$ $\bar{P}$-a.s. Thus

$$\limsup \frac{1}{n} \log f(\mathbf{X}_{k(n),n}) \le D \quad \bar{P}\text{-a.s.} \tag{5.5}$$

Now the event in (5.5) is in the tail sigma field, since $k(n) \to \infty$ and D is invariant. But P is absolutely continuous with respect to $\bar{P}$ on the tail sigma field. Thus the inequality in (5.5) also holds P-a.s. and using (5.3) we obtain

$$\limsup \frac{1}{n} \log f(\mathbf{X}_{1,n}) \le D \quad P\text{-a.s.} \tag{5.6}$$

It remains to show that $\liminf(1/n)\log f(\mathbf{X}_{1,n}) \ge D$.

By asymptotic dominance, given $\epsilon > 0$ there exists k such that $\lim(1/n)\log f(\mathbf{X}_{k,n}) = D$ except in a set of P probability less than ϵ. From Markov's inequality and the Borel–Cantelli lemma we find that $\liminf(1/n)\log f(\mathbf{X}_{k,n})/\bar{f}(\mathbf{X}_{k,n}) \ge 0$ P-a.s. Therefore,

$$\liminf \frac{1}{n} \log f(\mathbf{X}_{k,n}) \ge D \tag{5.7}$$

except in a set of P probability less than ϵ. But $\liminf(1/n)\log f(\mathbf{X}_{1,n})/f(\mathbf{X}_{k,n}) \ge 0$ P-a.s. [Compare with (5.3); here we only need that information densities are dominated from below.] Therefore,

$$\liminf \frac{1}{n} \log f(\mathbf{X}_{1,n}) \ge D \tag{5.8}$$

except in a set of P probability less than ϵ. Letting $\epsilon \to 0$, we find that (5.8) holds P-almost surely. Together (5.6) and (5.8) show a.s. convergence of the sequence of relative entropy densities for the asymptotically mean stationary process. This completes the proof of Theorem 3. □

6. Convergence of information densities. Let $\{(U_n, V_n)\}$ be a jointly stationary stochastic process. The random variables U_n and V_n take values in standard Borel spaces $(\mathbb{U}, \mathscr{B}_{\mathbb{U}})$ and $(\mathbb{V}, \mathscr{B}_{\mathbb{V}})$. For information theory applications we call elements of $\mathbb{U}$ and $\mathbb{V}$ input and output symbols, respectively. A central issue in information theory is the asymptotic behavior of the mutual information densities $(1/n)i(\mathbf{U}_{1,n}; \mathbf{V}_{1,n})$. As an important example we mention channel coding theory. If the sequence of information densities $(1/n)i(\mathbf{U}_{1,n}; \mathbf{V}_{1,n})$ converges in probability to a constant limit R, say, then by Shannon's random coding technique there exists a channel code of rate R: that is, there exists a sequence of codebooks containing e^{nR} different codewords (length n sequences of input symbols), such that if a codeword $\mathbf{U}_{1,n}$ is chosen (according to the uniform distribution over the e^{nR} codewords) and if the output $\mathbf{V}_{1,n}$ is conditionally

distributed according to $P_{\mathbf{V}_{1,n}|\mathbf{U}_{1,n}}$ (the channel) then the codeword $\mathbf{U}_{1,n}$ can be recovered by a measurable function of the output with probability of error tending to zero as $n \to \infty$.

For discrete stationary processes with finite cardinality for $\mathbb{U}$ and $\mathbb{V}$, the Shannon–McMillan–Breiman theorem readily yields the almost sure and L^1 convergence of the sequence of information densities $(1/n)i(\mathbf{U}_{1,n}; \mathbf{V}_{1,n})$. Pinsker [(1964), Theorem 8.2.1] and Gray and Kieffer [(1980b), Theorem 5] used a discretization approach to prove L^1 convergence for stationary processes satisfying a general condition: namely, that the information rate $\lim(1/n)I(\mathbf{U}_{1,n}; \mathbf{V}_{1,n})$ exists, is finite, and equals the Pinsker information rate I^* given by the supremum (over all finite partitions of $\mathbb{U}$ and $\mathbb{V}$) of the discrete information rates.

We obtain almost sure convergence of information densities for stationary ergodic processes. We remark that our conditions imply that the information rate exists and equals the Pinsker information rate, but do not imply that this rate is finite [see Pinsker (1964), Theorem 7.4.2].

Ergodicity is assumed only for simplicity. The results are easily extended to the nonergodic case as in Section 4 (if we assume finite information rate).

THEOREM 4. *Let $\{(U_n, V_n)\}$ be a jointly stationary ergodic process. If $I(U_0; \mathbf{U}_{-\infty,-m-1}|\mathbf{U}_{-m,-1})$ and $I(V_0; \mathbf{V}_{-\infty,-m-1}|\mathbf{V}_{-m,-1})$ are finite for some $m \geq 0$, then the sequence of mutual information densities $(1/n)i(\mathbf{U}_{1,n}; \mathbf{V}_{1,n})$ converges almost surely.*

PROOF. Set $X_n = (U_n, V_n)$. Suppose the information density $i(\mathbf{U}_{1,n}; \mathbf{V}_{1,n})$ is a.s. finite for all n. [Otherwise convergence is trivial, because if $i(\mathbf{U}_{1,n}; \mathbf{V}_{1,n})$ is infinite for some n then it is infinite for all larger n.] The mutual information densities may be expanded as

$$\frac{1}{n}i(\mathbf{U}_{1,n}; \mathbf{V}_{1,n}) = \frac{1}{n}\log \mathrm{f}(\mathbf{X}_{1,n}) - \frac{1}{n}\log \mathrm{f}(\mathbf{U}_{1,n}) - \frac{1}{n}\log \mathrm{f}(\mathbf{V}_{1,n}), \tag{6.1}$$

where $\mathrm{f}(\mathbf{U}_{1,n})$, $\mathrm{f}(\mathbf{V}_{1,n})$, and $\mathrm{f}(\mathbf{X}_{1,n})$ are the densities of $P_{\mathbf{U}_{1,n}}$, $P_{\mathbf{V}_{1,n}}$, and $P_{\mathbf{X}_{1,n}}$ with respect to dominating Markov measures $M_{\mathbf{U}_{1,n}}$, $M_{\mathbf{V}_{1,n}}$, and $M_{\mathbf{U}_{1,n}} \times M_{\mathbf{V}_{1,n}}$. In particular $M_{\mathbf{U}_{1,n}}$ can be constructed from the initial distribution $P_{\mathbf{U}_{1,m}}$ and the stationary transition probability $P_{U_{m+1}|\mathbf{U}_{1,m}}$. Similarly construct $M_{\mathbf{V}_{1,n}}$. The a.s. convergence of the three terms in (6.1) follows by Theorem 1. The limit of $(1/n)\log \mathrm{f}(\mathbf{U}_{1,n})$ is $I(U_0; \mathbf{U}_{-\infty,-m-1}|\mathbf{U}_{-m,-1})$ which is finite by assumption; likewise for $(1/n)\log \mathrm{f}(\mathbf{V}_{1,n})$. Since the limits of the last two terms of (6.1) are finite, the sequence of information densities $(1/n)i(\mathbf{U}_{1,n}; \mathbf{V}_{1,n})$ also converges. This completes the proof of Theorem 4. □

For most information theory applications, the channel is causal. By definition a causal channel $P_{\mathbf{V}_{1,\infty}|\mathbf{U}_{1,\infty}}$ satisfies $P_{\mathbf{V}_{1,n}|\mathbf{U}_{1,\infty}} = P_{\mathbf{V}_{1,n}|\mathbf{U}_{1,n}}$ for all n. Thus for a causal channel the present output V_n depends on past outputs $V_{1,n-1}$ and past and present inputs $\mathbf{U}_{1,n}$ but (conditionally) does not depend on future inputs. For such channels we obtain a.s. convergence of information densities under more general conditions.

THEOREM 5. *Let $\{(U_n, V_n)\}$ be a jointly stationary ergodic process with a causal channel $P_{\mathbf{V}_{1,n}|\mathbf{U}_{1,n}}$. If $I(V_0; \mathbf{V}_{-\infty,-m-1}|\mathbf{V}_{-m,-1})$ is finite for some $m \geq 0$, then the sequence of mutual information densities $(1/n)i(\mathbf{V}_{1,n})$ converges almost surely to the mutual information rate $I(\mathbf{U}_{-\infty,0}; V_0|\mathbf{V}_{-\infty,-1})$.*

PROOF. The information densities satisfy the chain rule $(i(\mathbf{U}_{1,n}; \mathbf{V}_{1,n}) = \sum_{j=0}^{n-1} i(\mathbf{U}_{1,n}; V_{j+1}|\mathbf{V}_{1,j})$. For a causal channel, the summands reduce to $i(V_{j+1}|\mathbf{V}_{1,j})$. So by stationarity we obtain

$$\frac{1}{n} i(\mathbf{U}_{1,n}; \mathbf{V}_{1,n}) = \frac{1}{n} \sum_{j=0}^{n-1} i(\mathbf{U}_{-j,0}; V_0|\mathbf{V}_{-j,-1}) \circ T^{j+1}. \tag{6.2}$$

By the chain rule we obtain the expansion

$$\begin{aligned} i(\mathbf{U}_{-n,0}; V_0|\mathbf{V}_{-n,-1}) = {} & i(V_0; \mathbf{U}_{-n,0}, \mathbf{V}_{-n,-m-1}|\mathbf{V}_{-m,-1}) \\ & - i(V_0; \mathbf{V}_{-n,-m-1}|\mathbf{V}_{-m,-1}). \end{aligned} \tag{6.3}$$

By Lemma 3, these terms converge a.s. and are L^1 dominated whenever $I(\mathbf{U}_{-\infty,0}; V_0|\mathbf{V}_{-\infty,-1})$ is finite. Hence by Breiman's generalized ergodic theorem

$$\lim \frac{1}{n} i(\mathbf{U}_{1,n}; \mathbf{V}_{1,n}) = I(\mathbf{U}_{-\infty,0}; V_0|\mathbf{V}_{-\infty,-1}) \quad \text{a.s.} \tag{6.4}$$

To treat the case $I(\mathbf{U}_{-\infty,0}; V_0|\mathbf{V}_{-\infty,-1}) = \infty$, note that the sequence of information densities $i(\mathbf{U}_{-n,0}; V_0|\mathbf{V}_{-n,-1})$ is dominated from below [since by assumption $I(V_0; \mathbf{V}_{-\infty,-m-1}|\mathbf{V}_{-m,-1})$ is finite]. Consequently, $\liminf(1/n)i(\mathbf{U}_{1,n}; \mathbf{V}_{1,n}) \geq I(\mathbf{U}_{-\infty,0}; V_0|\mathbf{V}_{-\infty,-1})$ almost surely. Therefore, even if the mutual information rate is infinite, the sequence of information densities converges as in (6.4). This completes the proof of Theorem 5. □

In information theory, a common class of channels are the memoryless channels. A memoryless channel is a causal channel satisfying $P_{\mathbf{V}_{1,n}|\mathbf{U}_{1,n}} = \times_{j=1}^{n} P_{V_j|U_j}$, with the same conditional distribution $P_{V_n|U_n}$ for all n. Given an input U_n, the output V_n is conditionally independent of all other inputs and outputs. In particular $I(V_0; \mathbf{V}_{-\infty,-1}|U_0) = 0$ so by the chain rule $I(V_0; \mathbf{V}_{-\infty,-1}) \leq I(V_0; U_0, \mathbf{V}_{-\infty,-1}) = I(V_0; U_0)$. Hence if $I(U_0; V_0)$ is finite, Theorem 5 applies to show a.s. convergence of the information densities. Thus we have established the following result which extends the L^1 convergence obtained by Pinsker (1963) and Kieffer (1978).

COROLLARY. *Let a stationary ergodic process $\{U_n\}$ be input to a memoryless channel $P_{V_n|U_n}$. If the mutual information $I(U_0; V_0)$ is finite, then the sequence of mutual information densities $(1/n)i(\mathbf{U}_{1,n}; \mathbf{V}_{1,n})$ converges almost surely to $I(U_0; V_0|\mathbf{V}_{\infty,-1})$.*

Acknowledgments. The author acknowledges the helpful suggestions of Professors Tom Cover, Robert Gray, Imre Csiszár, and Paul Algoet. The referee is thanked for extensive simplification of this paper and for suggesting that convergence of information densities be examined.

REFERENCES

BAHADUR, R. R. (1971). *Some Limit Theorems in Statistics*. SIAM, Philadelphia.

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

BREIMAN, L. (1960). A correction to "the individual ergodic theorem of information theory." *Ann. Math. Statist.* **31** 809–810.

CARLESON, L. (1958). Two remarks on the basic theorem of information theory. *Math. Scand.* **6** 175–180.

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.

DOOB, J. L. (1953). *Stochastic Processes*. Wiley, New York.

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

GRAY, R. M. and KIEFFER, J. C. (1980b). Mutual information rate, distortion, and quantization in metric spaces. *IEEE Trans. Inform. Theory* **26** 412–422.

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

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

KIEFFER, J. C. (1976). Correction to: "a counterexample to Perez's generalization of the Shannon–McMillan theorem." *Ann Probab.* **4** 153–154.

KIEFFER, J. C. (1978). Block coding for an ergodic source relative to a zero–one valued fidelity criterion. *IEEE Trans. Inform. Theory* **24** 432–438.

MCMILLAN, B. (1953). The basic theorems of information theory. *Ann. Math. Statist.* **24** 196–219.

MOY, S. C. (1961). Generalizations of Shannon–McMillan theorem. *Pacific J. Math.* **11** 705–714.

PARTHASARATHY, K. R. (1964). A note on McMillan's theorem for countable alphabets. *Trans. Third Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes*, 541–543. Czechoslovak Academy of Sciences, Prague.

PARTHASARATHY, K. R. (1967). *Probability Measures on Metric Spaces*. Academic, New York.

PEREZ, A. (1964). Extensions of Shannon–McMillan's limit theorem to more general stochastic processes. *Trans. Third Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes*. 545–574. Czechoslovak Academy of Sciences, Prague.

PINSKER, M. S. (1964). *Information and Information Stability of Random Variables*. Translation by A. Feinstein. Holden-Day, San Francisco.

PINSKER, M. S. (1963). Sources of messages. *Problemy Peredači Informacii* **14** 5–20.

SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4** 10–26.

SHANNON, C. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27** 379–423, 623–656.

TULCEA, A. I. (1960). Contributions to information theory for abstract alphabets. *Ark. Mat.* **4** 235–247.

WALD, A. (1949). Note on the consistency of the maximum likelihood estimate, *Ann. Math. Statist.* **20** 595–601.

ANDREW R. BARRON
DIVISION OF STATISTICS, 221 ALTGELD HALL
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF ILLNOIS
1409 WEST GREEN STREET
URBANA, IL 61801