

CODING THEOREM FOR STATIONARY, ASYMPTOTICALLY MEMORYLESS, CONTINUOUS-TIME CHANNELS

By T. T. KADOTA AND A. D. WYNER

Bell Telephone Laboratories, Incorporated

We give a mathematical definition of stationary, asymptotically memoryless, continuous-time channels, and prove a coding theorem and its converse for such channels.

1. Introduction. Consider a continuous channel described by

$$(1) \quad y(t) = \int_{-\infty}^t h(t-s)x(s) ds + z(t),$$

where x and y are the input and the output of the channel, h a filter impulse-response, and z an additive noise. Suppose z is a zero-mean, stationary, Gaussian process. Observe that the probability distribution of y is same as that of z except for the mean, which is the integral term in (1). Hence the distribution of the output is invariant under time shift if the input is similarly shifted (Property 1). Next, under certain conditions on h , x and z the distribution of the present output $y(t)$ depends diminishingly little on the input of the remote past (Property 2). Third, the noise covariance must vanish as $t \rightarrow \infty$. Hence the output values at two different time-points are asymptotically independent as the time-points are separated further apart (Property 3). We call a channel with Property 1, *stationary*, with Property 2, *asymptotically input-memoryless*; with Property 3, *asymptotically output-memoryless*. We call a continuous-time channel with Properties 1–3, *stationary and asymptotically memoryless*.

Suppose we consider increments of the output, instead of its values, and their probability distribution for a fixed x . Then, we can define the *incremental version* of the stationary, asymptotically memoryless, continuous-time channel. Obviously, the incremental version is more general than the non-incremental. For example, if z in (1) is a process with stationary, independent increments, then the channel described by (1) is incrementally stationary, incrementally input- and output-memoryless, while it is nonstationary and has infinite output-memory, in general, and it obviously has infinite input-memory. Here the incremental stationarity and the output-memorylessness are the direct consequence of z having stationary, independent increments. The incremental input-memorylessness follows from the fact that $y(t) - y(s)$ depends on x only during (s, t) .

In the next section, we give general mathematical definitions of the stationary, asymptotically memoryless, continuous-time channel and its incremental version, and prove a coding theorem and its converse for such channels.

Received September 20, 1971; revised March 10, 1972.

AMS 1970 subject classifications. Primary 94A15; Secondary 60G10.

Key words and phrases. Coding theorem, continuous-time channels, asymptotically memoryless channels.

Previously, Wolfowitz (1964) treated a special case (the discrete-time and finite-alphabet version) of our channel, where the input and the output are sequences of integers taken from a finite set. He defined stationary, output-memoryless, asymptotically input-memoryless channels and proved the coding theorem and the converse for such channels. He then indicated how to extend the results to asymptotically memoryless channels. Pfaffelhuber (1971) which appeared after this paper was written, also studied the same special case. Although he gave an apparently more general definition of the asymptotically memoryless channel, he did not prove the converse, thus failing to obtain the capacity. Our definition is a generalization of Pfaffelhuber's to the continuous-time and continuous-alphabet case, where the input and the output are functions on the real line. The capacity we obtain incorporates a given input constraint and takes into account the intersymbol interference.

2. Mathematical definitions and coding theorem. Let X and Y be spaces of all real functions x and y on $(-\infty, \infty)$. Denote by $\mathcal{A}(s, t]$ the σ -field generated by the class of cylinder sets $\{x: a_i < x(t_i) \leq b_i, i = 1, \dots, n\}$, $s < t_i \leq t$, where a_i, b_i, t_i and n are arbitrary. Denote by $\mathcal{B}(s, t]$ the σ -field of y sets similarly generated. Let $\nu(\cdot; x)$ be a transitional measure defined on $\mathcal{B}(-\infty, \infty)$ for every $x \in X$. By definition, $\nu(B; \cdot)$ is $\mathcal{A}(-\infty, \infty)$ -measurable. We refer to ν as a channel, X and Y as the input and the output spaces respectively. A channel ν is *causal* if $B \in \mathcal{B}(-\infty, t]$ implies $\mathcal{A}(-\infty, t]$ -measurability of $\nu(B; \cdot)$ for any B and t . ν is *stationary* if $\nu(S_u B; S_u x) = \nu(B, x)$ for any $B \in \mathcal{B}(-\infty, \infty)$, x and $u \in (-\infty, \infty)$, where S_u denotes the operation of shifting time by u . We say ν is *asymptotically input-memoryless* if for a given $\varepsilon > 0$ there exists $\tau(\varepsilon, s)$ such that $|\nu(B; x) - \nu_\tau(B; x)| < \varepsilon$ for any $B \in \mathcal{B}(s, \infty)$, x and s , where $\nu_\tau(B; \cdot)$, $B \in \mathcal{B}(s, \infty)$, is a $\mathcal{A}(s - \tau, \infty)$ -measurable version of $\nu(B; \cdot)$. In other words, $\nu(B; x_1)$ and $\nu(B; x_2)$ differ at most by ε for any x_1 and x_2 coinciding on $(s - \tau, \infty)$. If ν is stationary then τ is independent of s . This definition of asymptotic input-memorylessness on the whole input space X may be too restrictive. For proving the coding theorem, it is often sufficient to define it on a smaller space $X_1 \subset X$, which is usually specified by a given constraint on the input function. When such a definition is used we specify X_1 on which ν is asymptotically input-memoryless. We say ν is *asymptotically output-memoryless* if for a given $\varepsilon > 0$ there exists $\tau(\varepsilon; s, t)$ such that $B_1 \in \mathcal{B}(-\infty, s]$, $B_2 \in \mathcal{B}(t, \infty)$ and $t - s \geq \tau$ imply $|1 - \nu(B_1 B_2; x) / \nu(B_1; x) \nu(B_2; x)| < \varepsilon$ for any x, B_1 , and B_2 . This is similar to the strong mixing condition. If ν is stationary then τ is independent of s and t . We say ν is asymptotically memoryless if it is asymptotically input- and output-memoryless.

Let $\hat{\mathcal{B}}(s, t]$ be the σ -field generated by the class of all cylinder sets of the form $\{y: a_i < y(t_i) - y(s) \leq b_i, i = 1, \dots, n\}$, $s < t_i \leq t$, where a_i, b_i, t_i and n are arbitrary. Then we can define *incremental stationarity*, *incremental asymptotic input-memorylessness* and *incremental asymptotic output-memorylessness* by simply placing " $\hat{}$ " on the symbol \mathcal{B} in the preceding definitions.

Let P_X be a probability measure defined on $\mathcal{A}(-\infty, \infty)$. Define P_{XY} on $\mathcal{A}(-\infty, \infty) \times \mathcal{B}(-\infty, \infty)$ by $P_{XY}(A \times B) = \int_A \nu(B; x)P_X(dx)$, P_Y on $\mathcal{B}(-\infty, \infty)$ by $P_Y(B) = P_{XY}(X \times B)$, and $P_{X \times Y}$ on $\mathcal{A}(-\infty, \infty) \times \mathcal{B}(-\infty, \infty)$ by $P_{X \times Y}(A \times B) = P_X(A)P_Y(B)$, where $A \in \mathcal{A}(-\infty, \infty)$ and $B \in \mathcal{B}(-\infty, \infty)$ are arbitrary [7]. Let $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}}$ be sub σ -fields of $\mathcal{A}(-\infty, \infty)$ and $\mathcal{B}(-\infty, \infty)$ respectively, and $P_{XY|\tilde{\mathcal{A}} \times \tilde{\mathcal{B}}}$ and $P_{X \times Y|\tilde{\mathcal{A}} \times \tilde{\mathcal{B}}}$ the restrictions of P_{XY} and $P_{X \times Y}$ respectively to $\tilde{\mathcal{A}} \times \tilde{\mathcal{B}}$. Then the mutual information between the input and the output on $\tilde{\mathcal{A}} \times \tilde{\mathcal{B}}$ is given by

$$I_{\tilde{\mathcal{A}} \times \tilde{\mathcal{B}}} = \int_{X \times Y} \log f_{\tilde{\mathcal{A}} \times \tilde{\mathcal{B}}}(x, y)P_{XY}(d(x, y))$$

$$\text{if } P_{XY|\tilde{\mathcal{A}} \times \tilde{\mathcal{B}}} \text{ is absolutely continuous with respect to } P_{X \times Y|\tilde{\mathcal{A}} \times \tilde{\mathcal{B}}},$$

$$= \infty \quad \text{otherwise,}$$

where $f_{\tilde{\mathcal{A}} \times \tilde{\mathcal{B}}}$ denotes the Radon-Nikodym derivative of $P_{XY|\tilde{\mathcal{A}} \times \tilde{\mathcal{B}}}$ with respect to $P_{X \times Y|\tilde{\mathcal{A}} \times \tilde{\mathcal{B}}}$.

Denote by \mathcal{P}_T the class of P_X such that (i) $P_X(\bigcap_{i=-n}^n A_i) = \prod_{i=-n}^n P_X(A_i)$ for any $A_i \in \tilde{\mathcal{A}}(iT, (i+1)T]$, $i = 0, \pm 1, \dots, \pm n$, and any n , (ii) $P_X(S_{\pm nT} A_0) = P_X(A_0)$ for any n and $A_0 \in \tilde{\mathcal{A}}(0, T]$, (iii) $\Phi_T(x) \leq 1$ for almost every x with respect to P_X where $\tilde{\mathcal{A}}$ denotes a standard extension of \mathcal{A} (see the Remarks after the theorems) and Φ_T is a nonnegative $\tilde{\mathcal{A}}(0, T]$ -measurable functional. In other words, \mathcal{P}_T is the class of periodic and periodically independent measures with period T which satisfy the “ Φ_T -constraint”.

CODING THEOREM. Let ν be a casual, stationary, asymptotically memoryless, continuous-time channel. Put

$$C = \limsup_{T \rightarrow \infty} \sup_{P_X \in \mathcal{P}_T} \frac{1}{T} I_{\tilde{\mathcal{A}}(0, T]|\tilde{\mathcal{B}}(0, T]}.$$

For any $\epsilon > 0$ and $R < C$, there exist a set of $\{x_i(t), 0 < t \leq b\}$, $i = 1, \dots, m$, satisfying

$$(2) \quad \Phi_b(x_i) \leq 1, \quad i = 1, \dots, m; \quad b > 0,$$

and a $\mathcal{B}(0, b]$ -measurable m -partition (B_1, \dots, B_m) of Y such that

$$(3) \quad \sup_{\{x_i(t), -\infty < t \leq 0\}} \nu(B_i^c; x_i) \leq \epsilon, \quad i = 1, \dots, m.$$

for some b and m satisfying

$$(4) \quad (\log m)/b \geq R,$$

where B_i^c denotes the complement of B_i .

In the case where ν is only incrementally stationary and asymptotically memoryless, the assertion holds if “ $\hat{}$ ” is placed on the symbol \mathcal{B} .

CONVERSE THEOREM. For any $\{x_i(t), 0 < t \leq b\}$, $i = 1, \dots, m$, satisfying (2) and any $\mathcal{B}(0, b]$ -measurable m -partition (B_1, \dots, B_m) of Y , if

$$(\log m)/b \geq R > C,$$

then there exists $\epsilon(C, R) > 0$ such that

$$\nu(B_i^c; x_i) > \epsilon$$

for some $\{x_i(t), -\infty < t \leq 0\}$ for some $i, 1 \leq i \leq m$.

REMARKS. $\{x_i(t), 0 < t \leq b\}, i = 1, \dots, m$, represent code words of length b , and (2) is the Φ_b -constraint on them. The $\mathcal{B}(0, b]$ -measurable partition (B_1, \dots, B_m) represents the decoding rule (based on observing y during $(0, b]$) corresponding to the code $\{x_i(t), 0 < t \leq b, i = 1, \dots, m\}$, in the sense that $\{y(t), 0 < t \leq b\}$ is decoded as $\{x_i(t), 0 < t \leq b\}$ if $y \in B_i$. Thus $\nu(B_i^c; x_i)$ is the probability of decoding error corresponding to x_i , and $\sup_{\{x_i(t), -\infty < t \leq 0\}} \nu(B_i^c; x_i)$ is the least upper bound over all the possible code words used previously. R signifies a lower bound on the rate of information transmission. Thus, the coding theorem states that for any R less than C there are a code and its decoding rule such that, regardless of all the code words used previously, the decoding-error probability for each code word can be made arbitrarily small while maintaining the rate $(\log m)/b$ above the given level R . If the channel ν is asymptotically input-memoryless only on a set X_1 of all x such that $\Phi_T(x) \leq 1$ for any T , then (3) must be modified as follows:

$$\sup_{\{x_i(t), -\infty < t \leq 0; x_i \in X_1\}} \nu(B_i^c; x_i) < \epsilon, \quad i = 1, \dots, m.$$

Two typical examples of Φ_T -constraint are (i) the average-power constraint where $\Phi_T(x) = T^{-1} \int_0^T x^2(t) dt$ and (ii) the peak-power constraint where $\Phi_T(x) = \sup_{0 \leq t \leq T} x(t)$. In both cases (i) and (ii), Φ_T is not $\mathcal{A}(0, T]$ -measurable. For this reason we have introduced the standard extension $\tilde{\mathcal{A}}(0, T]$ of $\mathcal{A}(0, T]$. Namely, $\tilde{\mathcal{A}}(0, T]$ is the σ -field generated by the class of sets A of the form $A = A_1 X_1 + A_2(X - X_1)$ where A_1 and A_2 are in $\mathcal{A}(0, T]$ and $X_1 \subset X$ ([2], [1]). In case (i) $X_1 = \{x: \int_0^T x^2(t) dt \leq T\}$ and in case (ii) $X_1 = \{x: \sup_{0 \leq t \leq T} x(t) \leq 1\}$. Depending on the choice of Φ_T , C may be finite or infinite. If it is infinite, there is no limitation on the rate in order to achieve an arbitrarily small decoding-error probability.

On the other hand, the converse theorem states that if R is greater than C then there is no code and the corresponding rule such that the decoding-error probability for each code word can be made arbitrarily small regardless of all the preceding code words, while maintaining the rate $(\log m)/b$ above the given level R . In fact we prove a stronger version. Namely, if $R > C$ then there is no code and the decoding-rule such that the decoding-error probability averaged over all the preceding code words can be made arbitrarily small, regardless of the probability measure for the preceding code words which belongs to the class \mathcal{P}_b , while maintaining the rate above the given level.

As a result of the coding theorem and the converse, C is the *capacity* of the causal, stationary, asymptotically memoryless, continuous-time channel with the Φ_T -constraint. It should be noted that this capacity takes into account the effect of intersymbol interference from the preceding code words.

3. Proof of Coding Theorem.

(a) Suppose $C < \infty$. From the definition of C , $R < C$ implies the existence of some T and $P_X \in \mathcal{P}_T$ such that

$$(5) \quad R < \frac{1}{T} I_{\tilde{\mathcal{V}}(0,T) \cdot \mathcal{B}(0,T)}$$

Using such a T , define

$$\begin{aligned} \mathcal{A}_i &= \tilde{\mathcal{A}}(iT, (i + 1)T), & \mathcal{B}_i &= \mathcal{B}(iT, (i + 1)T), & i &= 0, 1, \dots, \\ \mathcal{V}^{(n)} &= \bigvee_{i=0}^{n-1} \mathcal{A}_i, & \mathcal{B}^{(n)} &= \bigvee_{i=0}^{n-1} \mathcal{B}_i, & n &= 1, 2, \dots \end{aligned}$$

We prove in Appendix I that P_{XY} is strongly mixing relative to $\{\mathcal{A}_i \times \mathcal{B}_i\}$, and in Appendix II that $n^{-1}I_{\mathcal{V}^{(n)} \cdot \mathcal{B}^{(n)}}$ is a nondecreasing function of n . Then, from the definition of C and the assumption $C < \infty$, $\lim_{n \rightarrow \infty} n^{-1}I_{\mathcal{V}^{(n)} \cdot \mathcal{B}^{(n)}}$ exists and is finite, and we denote it by \bar{I} . The strong mixing of P_{XY} and $\bar{I} < \infty$ satisfy the conditions of the information-stability theorem in [9], page 117, which asserts that $n^{-1} \log f_{\mathcal{V}^{(n)} \times \mathcal{B}^{(n)}}$ converges to \bar{I} in probability with respect to P_{XY} . Then we can choose m and n which simultaneously satisfy

$$(6) \quad \frac{1}{nT} \log m \geq R,$$

$$(7) \quad \exp[-n(\bar{I} - \delta) + \log m] < \frac{\varepsilon}{4},$$

$$(8) \quad P_{XY} \left(\left\{ (x, y) : \frac{1}{n} \log f_{\mathcal{V}^{(n)} \times \mathcal{B}^{(n)}}(x, y) < \bar{I} - \delta \right\} \right) < \frac{\varepsilon}{4}$$

for some $\delta > 0$. For example, take $\delta = (\bar{I} - RT)/3$ and choose n sufficiently large so that $n\delta > 1$ and $\exp(-n\delta) < \varepsilon/4$ and (8) holds. Note $\delta > 0$ because $n^{-1}I_{\mathcal{V}^{(n)} \cdot \mathcal{B}^{(n)}}$ is non-decreasing and $I_{\tilde{\mathcal{V}}(0,T) \cdot \mathcal{B}(0,T)} > RT$ from (5). With such a choice of δ and n , the left-hand side of (7) becomes $\exp(-nRT - 2\delta n + \log m)$, which can be made less than $\exp(-n\delta)$ by taking m such that $nRT \leq \log m < nRT + n\delta$. This choice of m obviously satisfies (6). Now, with (7) and (8), Feinstein's lemma [5] asserts the existence of $\tilde{x}_1, \dots, \tilde{x}_m$ and a $\mathcal{B}(0, nT)$ -measurable m -partition $(\hat{B}_1, \dots, \hat{B}_m)$ of Y such that $\nu(\hat{B}_i^c; \tilde{x}_i) < \varepsilon/2, i = 1, \dots, m$. $\tilde{x}_1, \dots, \tilde{x}_m$ can be chosen to satisfy

$$\Phi_T(S_{\pm iT} x) < 1, \quad i = 1, \dots, m; \quad j = 0, \pm 1, \dots,$$

since $P_X \in \mathcal{P}_T$. Because ν is causal and asymptotically input-memoryless, there exists an integer $l > 0$ such that $|\nu(\hat{B}_i^c; \tilde{x}_i) - \nu(\hat{B}_i^c; \hat{x}_i)| < \varepsilon/2$ for any \hat{x}_i which satisfies $\hat{x}_i(t) = \tilde{x}_i(t), -lT < t \leq nT$. Hence, for such $\hat{x}_1, \dots, \hat{x}_m, \nu(\hat{B}_i^c; \hat{x}_i) < \varepsilon, i = 1, \dots, m$. Then put $x_i = S_{iT} \hat{x}_i$ and $B_i = S_{iT} \hat{B}_i, i = 1, \dots, m$, and $b = (l + n)T$. This proves the theorem.

(b) Suppose $C = \infty$. From the definitions of C and asymptotic output-memorylessness, for any $R < \infty$ and $\varepsilon > 0$ there exist some T and $P_X \in \mathcal{P}_T$

such that

$$R < \frac{1}{2T} I_{\mathcal{A}(0, T], \mathcal{B}(0, T)},$$

$$(9) \quad |1 - \nu(B_1 B_2; X) / \nu(B_1; X) \nu(B_2; X)| < \varepsilon$$

for any $B_1 \in \mathcal{B}(-\infty, t]$, $B_2 \in \mathcal{B}(t + T, \infty)$ and an arbitrary t . Substituting (9) into (12) in Appendix I and following the deduction thereafter, we can establish that

$$(10) \quad |1 - P_{XY}(\Gamma \Lambda) / P_{XY}(\Gamma) P_{XY}(\Lambda)| < \varepsilon$$

for any $\Gamma \in \mathcal{A}(-\infty, t] \times \mathcal{B}(-\infty, t]$, $\Lambda \in \mathcal{A}(t + T, \infty) \times \mathcal{B}(t + T, \infty)$. Now with the above choice of T , define

$$\begin{aligned} \tilde{\mathcal{A}}_i &= \mathcal{A}(2iT, (2i + 1)T], & \tilde{\mathcal{B}}_i &= \mathcal{B}(2iT, (2i + 1)T], & i &= 0, 1, \dots, \\ \tilde{\mathcal{A}}^{(n)} &= \bigvee_{i=0}^{n-1} \tilde{\mathcal{A}}_i, & \tilde{\mathcal{B}}^{(n)} &= \bigvee_{i=0}^{n-1} \tilde{\mathcal{B}}_i, & n &= 0, 1, \dots. \end{aligned}$$

Observe that the proofs of Appendices I and II can be used to establish that P_{XY} is strongly mixing relative to $\{\tilde{\mathcal{A}}_i \times \tilde{\mathcal{B}}_i\}$ and $n^{-1} I_{\tilde{\mathcal{A}}^{(n)}, \tilde{\mathcal{B}}^{(n)}}$ is a non-decreasing function of n . By using (10) we prove in Appendix III that $\lim_{n \rightarrow \infty} n^{-1} I_{\tilde{\mathcal{A}}^{(n)}, \tilde{\mathcal{B}}^{(n)}} < \infty$. The remainder of the proof is identical to the case (a), with T replaced by $2T$.

PROOF OF CONVERSE THEOREM. Let $x_i, i = 1, \dots, m$, be any m elements in X satisfying (2). Denote by $\tilde{\mathcal{P}}_b$ the class of $P_X \in \mathcal{P}_b$ such that $P_X(A)$ is equal to the reciprocal of the number of x_i 's belonging to A . Put

$$\hat{\nu}(B_i^c; x_i) = E[\nu(B_i^c; \cdot) | \tilde{\mathcal{A}}(0, b)]_{x=x_i},$$

where the conditional expectation is with respect to some $P_X \in \tilde{\mathcal{P}}_b$. Define

$$P_e = \frac{1}{m} \sum_{i=1}^m \hat{\nu}(B_i^c; x_i),$$

$$\hat{I}_{\tilde{\mathcal{A}}(0, b], \tilde{\mathcal{A}}(0, b]} = \sum_{i,j=1}^m \frac{1}{m} \hat{\nu}(B_j; x_i) \log \left[\hat{\nu}(B_j; x_i) / \sum_{k=1}^m \frac{1}{m} \hat{\nu}(B_j; x_k) \right],$$

$$\kappa_m(\rho) = \rho \log(m - 1) - \rho \log \rho - (1 - \rho) \log(1 - \rho), \quad 0 \leq \rho \leq 1.$$

Then, by following the proof in [3], pages 78–79, it is straightforward to show that

$$\kappa_m(P_e) \geq \log m' - \hat{I}_{\tilde{\mathcal{A}}(0, b], \tilde{\mathcal{A}}(0, b]}.$$

Observe that for any $P_X \in \tilde{\mathcal{P}}_b$ and any m -partition (A_1, \dots, A_m) of X such that $x_i \in A_i$ and $x_j \notin A_i, j \neq i$,

$$\begin{aligned} \hat{I}_{\tilde{\mathcal{A}}(0, b], \tilde{\mathcal{A}}(0, b]} &= \sum_{i,j=1}^m P_{XY}(A_i \times B_j) \log [P_{XY}(A_i \times B_j) / P_{X \times Y}(A_i \times B_j)] \\ &\leq I_{\tilde{\mathcal{A}}(0, b], \tilde{\mathcal{A}}(0, b]}. \end{aligned}$$

Hence, it follows from $\tilde{\mathcal{P}}_b \subset \mathcal{P}_b$ and the definition of C and Appendix II that

$$\hat{I}_{\tilde{\mathcal{A}}(0, b], \tilde{\mathcal{A}}(0, b]} \leq \sup_{P_X \in \tilde{\mathcal{P}}_b} I_{\tilde{\mathcal{A}}(0, b], \tilde{\mathcal{A}}(0, b]} \leq bC.$$

Then,

$$\kappa_m(P_e) \geq \log m - bC \geq (1 - C/R) \log m.$$

Now $\kappa_m(\rho)$ is a positive, concave function of ρ , $0 < \rho < 1$, vanishing at $\rho = 0$, $\max_{0 \leq \rho \leq 1} \kappa_m(\rho) = \log m$ at $\rho = (m - 1)/m$, $\kappa_m(1) = \log(m - 1)$ ([3], page 78). Define the inverse $\kappa_m^{-1}(\sigma)$, $0 \leq \sigma \leq \log m$, as the smaller of the two solutions ρ_1 and ρ_2 of $\kappa_m(\rho) = \sigma$. Then

$$P_\epsilon \geq \kappa_m^{-1}((1 - C/R) \log m).$$

Note $\kappa_m^{-1}((1 - C/R) \log m) > \kappa_m^{-1}((1 - C/R) \log(m - 1))$, and $\kappa_m^{-1}((1 - C/R) \times \log(m - 1))$ is in $(0, 1 - C/R)$ and is an increasing function of m , and $\kappa_m^{-1}(\sigma)$ is a decreasing function of m . Hence $P_\epsilon \geq \epsilon$ where $\epsilon = \kappa_3^{-1}((1 - C/R) \log 2)$. Then it follows from the definition of P_ϵ that $\nu(B_i^c; x_i) \geq \epsilon$ for some i , $1 \leq i \leq m$. Similarly, it follows from the definitions of $\hat{\nu}(B_i^c; x_i)$ and $\hat{\mathcal{F}}_b$ and from $P_X \in \hat{\mathcal{F}}_b$ that $\nu(B_i^c; x_i) \geq \epsilon$ for some $\{x_i(t), -\infty < t \leq 0\}$, for some i , $1 \leq i \leq m$.

APPENDICES

I. P_{XY} is strongly mixing relative to $\{\mathcal{A}_n \times \mathcal{B}_n\}$, i.e.,
 (11) $\lim_{n \rightarrow \infty} P_{XY}(\Gamma S_{nT} \Lambda) = P_{XY}(\Gamma)P_{XY}(\Lambda)$
 for any $\Gamma, \Lambda \in \mathcal{A}_0 \times \mathcal{B}_0$.

PROOF. For simplicity, put $A_n = S_{nT} A$, $B_n = S_{nT} B$, $n = 0, 1, \dots$, for any $A \in \mathcal{A}_0$, $B \in \mathcal{B}_0$. Then, from the definition of \mathcal{P}_T and the assumptions on ν , for any $P_X \in \mathcal{P}_T$ and any $A, A' \in \mathcal{A}_0$ and $B, B' \in \mathcal{B}_0$,

$$\begin{aligned} & |P_{XY}((A \times B)(A_n' \times B_n')) - P_{XY}(A \times B)P_{XY}(A_n' \times B_n')| \\ &= |\int_X \mathcal{L}_{AA_n'}(x)\nu(BB_n'; x)P_X(dx) \\ & \quad - P_{XY}(A \times B) \int_X \mathcal{L}_{A_n'}(x)\nu(B_n'; x)P_X(dx)| \\ (12) \quad & \leq \int_X \mathcal{L}_{AA_n'}(x)|\nu(BB_n'; x) - \nu(B; x)\nu(B_n'; x)|P_X(dx) \\ & \quad + \int_X \mathcal{L}_{AA_n'}(x)\nu(B; x)|\nu(B_n'; x) - \nu_{(n-1)T}(B_n'; x)|P_X(dx) \\ & \quad + P_{XY}(A \times B) \int_X \mathcal{L}_{A_n'}(x)|\nu_{(n-1)T}(B_n'; x) - \nu(B_n'; x)|P_X(dx), \end{aligned}$$

which vanishes as $n \rightarrow \infty$, where we have used

$$\begin{aligned} & \int_X \mathcal{L}_{AA_n'}(x)\nu(B; x)\nu_{(n-1)T}(B_n'; x)P_X(dx) \\ &= \int_X \mathcal{L}_A(x)\nu(B; x)P_X(dx) \int_X \mathcal{L}_{A_n'}(x)\nu_{(n-1)}(B_n'; x)P_X(dx) \end{aligned}$$

which follows from the definition of $\nu_{(n-1)T}$. Hence (11) holds when Γ and Λ are rectangles. Then, since any element of the field \mathcal{F}_0 of all rectangles $A \times B$, $A \in \mathcal{A}_0$, $B \in \mathcal{B}_0$, can be expressed as a finite union of disjoint rectangles, (11) holds for any $\Gamma, \Lambda \in \mathcal{F}_0$. Finally, that (11) holds for arbitrary $\Gamma, \Lambda \in \mathcal{A}_0 \times \mathcal{B}_0$, is seen from the fact that (i) given $\Gamma \in \mathcal{A}_0 \times \mathcal{B}_0$ and $\epsilon > 0$ there exists $\tilde{\Gamma} \in \mathcal{F}_0$ such that $P_{XY}(\Gamma \triangle \tilde{\Gamma}) < \epsilon$ ([4] page 56), and (ii)

$$\begin{aligned} & |P_{XY}(\Gamma S_{nT} \Lambda) - P_{XY}(\Gamma)P_{XY}(\Lambda)| \\ & \leq |P_{XY}(\Gamma S_{nT} \Lambda) - P_{XY}(\tilde{\Gamma} S_{nT} \tilde{\Lambda})| + |P_{XY}(\tilde{\Gamma} S_{nT} \tilde{\Lambda}) - P_{XY}(\tilde{\Gamma})P_{XY}(\tilde{\Lambda})| \\ & \quad + |P_{XY}(\tilde{\Gamma})P_{XY}(\tilde{\Lambda}) - P_{XY}(\Gamma)P_{XY}(\Lambda)|, \end{aligned}$$

where $\tilde{\Gamma}, \tilde{\Lambda} \in \mathcal{F}_0$.

II. $n^{-1}I_{\mathcal{A}_i(n), \mathcal{B}_i(n)}$ is non-decreasing.

PROOF. Since $I_{\mathcal{A}_i(n), \mathcal{B}_i(n)}$ is non-decreasing in n , it suffices to show that $I_{\mathcal{A}_i(n), \mathcal{B}_i(n)}$ is convex. For notational convenience, we write $I(X_k^l Y_k^l, Y_m^n)$ for $I_{[\mathcal{V}_{i=k+1}^l(\mathcal{A}_i \times \mathcal{B}_i)] \cdot (\mathcal{V}_{i=m+1}^n \mathcal{B}_i)}$, $I(X_k^l, Y_k^l | Y_m^n)$ for $I_{[\mathcal{V}_{i=k+1}^l(\mathcal{A}_i) \cdot (\mathcal{V}_{i=k+1}^l \mathcal{B}_i) | \mathcal{V}_{i=m+1}^n \mathcal{B}_i]}$, etc., where the second is the conditional information of $\mathcal{V}_{i=k+1}^l \mathcal{A}_i$ and $\mathcal{V}_{i=k+1}^l \mathcal{B}_i$ given $\mathcal{V}_{i=m+1}^n \mathcal{B}_i$ as defined in [6]. Then, with the use of Kolmogorov's identity [9], page 31; [6]

$$\begin{aligned} I(X_0^n, Y_0^n) &= I(X_0^{n-1}, Y_0^n) + I(X_{n-1}^n, Y_0^n | X_0^{n-1}) \\ &= I(X_0^{n-1}, Y_0^{n-1}) + I(X_0^{n-1}, X_{n-1}^n | Y_0^{n-1}) + I(X_{n-1}^n, X_0^{n-1} Y_0^n) \\ &= I(X_0^{n-1}, Y_0^{n-1}) + I(X_0^{n-1} Y_0^{n-1}, X_{n-1}^n Y_{n-1}^n) + I(X_{n-1}^n, Y_{n-1}^n) \\ &\quad - I(Y_0^{n-1}, Y_{n-1}^n) \end{aligned}$$

where $I(X_0^{n-1}, X_{n-1}^n) = 0$ has been used. Thus, using the stationarity of P_{XY} relative to $\{\mathcal{A}_i \times \mathcal{B}_i\}$,

$$(13) \quad \begin{aligned} I(X_0^n, Y_0^n) - I(X_0^{n-1}, Y_0^{n-1}) \\ = I(X_0^1, Y_0^1) + I(X_{-n+1}^0 Y_{-n+1}^0, X_0^1 Y_0^1) - I(Y_{-n+1}^0, Y_0^1). \end{aligned}$$

Hence

$$\begin{aligned} I(X_0^{n+1}, Y_0^{n+1}) - I(X_0^n, Y_0^n) &- [I(X_0^n, Y_0^n) - I(X_0^{n-1}, Y_0^{n-1})] \\ &= I(X_{-n}^0 Y_{-n}^0, X_0^1 Y_0^1) - I(X_{-n+1}^0 Y_{-n+1}^0, X_0^1 Y_0^1) - I(Y_{-n}^0, Y_0^1) \\ &\quad + I(Y_{-n+1}^0, Y_0^1) \\ &= I(X_{-n}^{-n+1} Y_{-n}^{-n+1}, X_0^1 Y_0^1 | X_{-n+1}^0 Y_{-n+1}^0) \\ &\quad - I(Y_{-n}^{-n+1}, Y_0^1 | Y_{-n+1}^0) \\ &\geq 0, \end{aligned}$$

where Kolmogorov's identity is used for the second equality and Jensen's inequality for the last inequality.

III. $\lim_{n \rightarrow \infty} n^{-1}I_{\tilde{\mathcal{A}}_i(n), \tilde{\mathcal{B}}_i(n)} < \infty$.

PROOF. According to the definition of mutual information,

$$I(X_{-\infty}^{-1} Y_{-\infty}^{-1}, X_0^1 Y_0^1) = \sup_{\{\Gamma_i, \Lambda_j\}} \sum_{i,j} P_{XY}(\Gamma_i \Lambda_j) \log \frac{P_{XY}(\Gamma_i \Lambda_j)}{P_{XY}(\Gamma_i) P_{XY}(\Lambda_j)}$$

where the supremum is taken over all rectangular partitions $\{\Gamma_i, \Lambda_j\}$ of $X \times Y$; $\Gamma_i \in \mathcal{A}(-\infty, -T] \times \mathcal{B}(-\infty, -T]$, $\Lambda_j \in \mathcal{A}(0, T] \times \mathcal{B}(0, T]$. By substituting (10),

$$I(X_{-\infty}^{-1} Y_{-\infty}^{-1}, X_0^1 Y_0^1) \leq \log(1 + \hat{\varepsilon}) < \infty.$$

Again, for notational convenience, we write $I(\tilde{X}_k^l \tilde{Y}_k^l, \tilde{Y}_m^n)$ for

$$I_{[\mathcal{V}_{i=k+1}^l(\tilde{\mathcal{A}}_i \times \tilde{\mathcal{B}}_i)] \cdot (\mathcal{V}_{i=m+1}^n \tilde{\mathcal{B}}_i)}, \quad \text{etc.}$$

Then, from (13),

$$\begin{aligned} \frac{1}{n} I(\tilde{X}_0^n, \tilde{Y}_0^n) &= \frac{1}{n} \sum_{i=2}^n [I(\tilde{X}_0^i, \tilde{Y}_0^i) - I(\tilde{X}_0^{i-1}, \tilde{Y}_0^{i-1})] + \frac{1}{n} I(\tilde{X}_0^1, \tilde{Y}_0^1) \\ &\leq I(\tilde{X}_0^1, \tilde{Y}_0^1) + \frac{1}{n} \sum_{i=2}^n I(\tilde{X}_{-i+1}^0, \tilde{Y}_{-i+1}^0, \tilde{X}_0^1, \tilde{Y}_0^1) \\ &\leq I(\tilde{X}_0^1, \tilde{Y}_0^1) + I(\tilde{X}_{-\infty}^0, \tilde{Y}_{-\infty}^0, \tilde{X}_0^1, \tilde{Y}_0^1). \end{aligned}$$

Since $n^{-1}I(\tilde{X}_0^n, \tilde{Y}_0^n)$ is non-decreasing, $\lim_{n \rightarrow \infty} n^{-1}I(\tilde{X}_0^n, \tilde{Y}_0^n)$ exists and is finite.

REFERENCES

- [1] BHARUCHA, B. H. (1969). A posteriori distributions and detection theory. *Information and Control* **14** 98-132.
- [2] DOOB, J. L. (1953). *Stochastic Processes*. Wiley, New York, 69.
- [3] GALLAGER, R. G. (1968). *Information Theory and Reliable Communication*. Wiley, New York.
- [4] HALMOS, P. R. (1950). *Measure Theory*. Van Nostrand, Princeton.
- [5] KADOTA, T. T. (1970). Generalization of Feinstein's fundamental lemma. *IEEE Trans. Information Theory* **16** 791-792.
- [6] KADOTA, T. T. (1971). On the mutual information of memoryless channels. *IEEE Trans. Information Theory* **17** 140-142.
- [7] NEVEU, J. (1965). *Mathematical Foundations of the Calculus of Probability*. Holden-Day, San Francisco, 73-77.
- [8] PFAFFELHUBER, E. (1971). Channels with asymptotically decreasing memory and anticipation. *IEEE Trans. Information Theory* **17** 379-385.
- [9] PINSKER, M. S. (1964). *Information and Information Stability of Random Variables and Processes*. Holden-Day, San Francisco.
- [10] WOLFOWITZ, J. (1964). *Coding Theorems of Information Theory*. Springer-Verlag, New York.

BELL TELEPHONE LABORATORIES
MURRAY HILL, NEW JERSEY 07974