INFORMATION AND CONTROL 55, 238–264 (1982)

Channel Distances and Representation*

D. L. NEUHOFF

Department of Electrical and Computer Engineering,
University of Michigan, Ann Arbor, Michigan 48109

AND

P. C. SHIELDS[†]

Department of Mathematics, Stanford University,
Stanford, California 94305

The properties of several distance measures for discrete stationary channels with memory are studied. All are based on Ornstein's $\bar{d}$-random process distance. The strongest of these distances has been employed in a theory concerned with the approximation of $\bar{d}$-continuous conditionally almost block independent (CABI) channels by primitive and other simple models. Here the approximation with respect to the weaker distances and the equivalence of the weaker distances to the strongest is investigated. In addition, an exact representation of a $\bar{d}$-continuous CABI channel as an infinite sliding-block coding of the input joined with an I.I.D. noise source is developed.

I. INTRODUCTION

In earlier work on the approximation of discrete stationary channels with memory by channels with finite structures (Neuhoff and Shields 1979, 1982a, and 1982b) we employed a measure of distance between channels to quantify the degree to which one channel approximates another. Specifically, we introduced a channel distance measure which is a generalization of Ornstein's $\bar{d}$-distance for random processes (Ornstein 1973) and which we denote here by $\bar{D}$. We showed that the class of channels that can be arbitrarily well approximated in $\bar{D}$ by either finite memory, primitive, or indecomposable finite state channels is characterized by the properties of $\bar{d}$-continuity, an input memory decay condition due to Gray and Ornstein (1973), and conditional almost block independence (CABI), an output

* Research sponsored by AFOSR Grant AFOSR-80-0054. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes.

[†] On leave from the Mathematics Department, University of Toledo, Toledo, Ohio 43606.

0019-9958/82 $2.00

memory decay condition (Neuhoff and Shields 1979). Clearly the significance of this work rests on the suitability of the $\bar{D}$-distance. In this paper we explore its suitability by comparing it with several other candidates for distance measures. In addition as part of one aspect of the comparisons we develop an exact representation of a $\bar{d}$-continuous, CABI channel as an infinite sliding-block coding of the input joined with an I.I.D. noise source. This result is of interest in its own right and hence is discussed in a separate section.

Ideally, the distance measure one chooses should be a metric or pseudometric and should be the weakest distance having the property that if channels are close in this distance, then channel capacities are close and the performances obtained using a fixed channel code are also close. The $\bar{D}$-distance is a pseudometric and it has the required capacity and performance continuity. Furthermore if two channels have $\bar{D}$-distance 0, then they are *equivalent* in the sense that for any block-stationary input process the input–output pair processes are identical. From the communications point of view equivalent channels are indistinguishable. On the other hand, the $\bar{D}$-distance is so strong that it assigns nonzero distance to some pairs of equivalent channels. In this paper we consider several weaker distances. Among them are two that have the properties that channels are equivalent if and only if they are zero distance apart and that close channels have similar capacities and code performances. In addition we show that for the important class of $\bar{d}$-continuous channels our original measure $\bar{D}$ is uniformly equivalent to the weakest distance we consider.

An outline of the paper follows. Section II contains notation and definitions. Section III contains definitions of various channel distances and statements of their properties. In Section IV we discuss channel approximations and in Section V the completeness and exact representation of the $\bar{d}$-continuous, CABI channels. Section VI has proofs of the results of Section III. Finally, the Appendix lists a number of useful properties we will draw upon.

II. Preliminaries

If A_1 is a set and $\mathbf{A}_1$ a σ-algebra of subsets of A_1, then A_m^n denotes the set of all sequences $X_m^n = (x_m, \ldots, x_n)$ with $x_i \in A_1$ and $\mathbf{A}_m^n$ denotes the usual product σ-algebra for A_m^n. If A_1 is a finite set, we take $\mathbf{A}_1$ to be the set of all subsets of A_1. We shall usually write x^n, A^n, and $\mathbf{A}^n$ instead of x_1^n, A_1^n, and $\mathbf{A}_1^n$, and x, A, and $\mathbf{A}$ instead of $x_{-\infty}^{\infty}$, $A_{-\infty}^{\infty}$, and $\mathbf{A}_{-\infty}^{\infty}$. If $m \leqslant k \leqslant l \leqslant n$, then $\langle a_k^l \rangle$ denotes the cylinder set in A_m^n determined by a_k^l, that is, the set of all x_m^n such that $x_i = a_i$, $k \leqslant i \leqslant l$. The coordinate function shall be denoted by X_i, that is, $X_i(x_m^n) = x_i$, $m \leqslant i \leqslant n$.

If α is a probability measure on A_m^n, that is, on the measurable space $(A_m^n, \mathbf{A}_m^n)$, and $m \leqslant k \leqslant l \leqslant n$, then α_k^l denotes the measure induced by α on A_k^l. We will often write α^l instead of α_1^l and write $\alpha(a_k^l)$ in place of $\alpha(\{a_k^l\})$ or $\alpha(\langle a_k^l \rangle)$. The coordinate functions $(X_m,..., X_n)$ on A_m^n are random variables with distribution governed by α. We usually write X^n in place of X_1^n. When $m = -\infty$ and $n = +\infty$, the sequence of random variables $X_{-\infty}^{\infty}$ is called a random process, or simply a process, and is ordinarily denoted by X or α.

Let N be a positive integer. A process α is called *N-stationary* if $\alpha(T^N E) = \alpha(E)$, all $E \in \mathbf{A}$, where T denotes the shift operation: $(Tx)_i = x_{i+1}$. A process is *stationary* if it is 1-stationary and *block stationary* if it is N-stationary for some N. A stationary process is *ergodic* if for any invariant set E, $\alpha(E)$ is 0 or 1.

The *ergodic decomposition* $[\beta_\theta, w]$ of a stationary process α is described as follows (Gray and Davisson 1974, Rohlin 1962) there is a probability space (Θ, Γ, w) such that for each $\theta \in \Theta$ there is an ergodic source α_θ such that for each $E \in \mathbf{A}$, $\theta \to \alpha_\theta(E)$ is Γ measurable and

$$\alpha(E) = \int \alpha_\theta(E)\, dw(\theta).$$

Given two measures α and β on A_m^n, $\alpha \nabla \beta$ denotes the set of all measures ω on $A_m^n \times A_m^n$ with α and β as marginals. Any such ω is called a *joining* of α and β.

Given two sequences $x^n, y^n \in A^n$, the normalized Hamming distance $d_n(x^n, y^n)$ between them is defined to be the number of places in which they disagree divided by n. For infinite sequences $x, y \in A$ we define $d_n(x, y) = d_n(x^n, y^n)$ and $d(x, y) = \limsup_{n \to \infty} d_n(x, y)$.

The $\bar{d}_n$ distance (Ornstein 1973) between two measures α and β on A^n, $n < \infty$, is defined by

$$\bar{d}_n(\alpha, \beta) = \inf_{\omega \in \alpha \nabla \beta} E_\omega[d_n(X^n, Y^n)],$$

where X^n and Y^n are the coordinate random variables associated with α and β, respectively, and $E_\omega[\ ]$ denotes expectation with respect to the measure ω.

The $\bar{d}$ distance between processes α and β is defined by

$$\bar{d}(\alpha, \beta) = \limsup_{n \to \infty} \bar{d}_n(\alpha, \beta).$$

If α and β are stationary or block-stationary, then the lim sup above is in fact a limit. Several other important properties of $\bar{d}_n$ and $\bar{d}$ are listed in the Appendix.

A channel $[A, B, \nu, X, Y]$ is characterized by an input alphabet A_1, an output alphabet B_1 and a family of measures $\{\nu_x : x \in A\}$ on the output space

B such that for any $F \in \mathbf{B}$, $\nu_x(F)$ is an $\mathbf{A}$-measurable function of x. The channel input and output at time n are labelled X_n and Y_n, respectively, and the sequences of inputs and outputs are labelled X and Y. When $X = x$, then Y is a random process characterized by the measure ν_x. Such a channel will ordinarily be denoted ν. We will restrict attention to channels that have finite input and output alphabets and that are stationary; i.e.

$$\nu_{Tx}(TF) = \nu_x(F), \qquad x \in A, \quad F \in \mathbf{B}.$$

In general ν_x is a non-stationary process, however, if ν is stationary and x is periodic with period N, then ν_x is N-stationary.

A source is a random process α with alphabet the same as the channel's input alphabet A_1. By $\alpha\nu$ we mean the pair process X, Y that results when the source $\alpha = X$ is the input to the channel ν. The measure $\alpha\nu$ is specified by

$$\alpha\nu(E \times F) = \int_E \nu_x(F)\, d\alpha(x), \qquad E \in \mathbf{A}, \quad F \in \mathbf{B}.$$

If α and ν are each stationary (N-stationary), then $\alpha\nu$ is also stationary (N-stationary).

We now describe several categories of channels. A channel ν is *memoryless* if for each $a \in A_1$ there exists a measure μ_a on B_1 such that for any cylinder set $\langle b_m^n \rangle$,

$$\nu_x(\langle b_m^n \rangle) = \prod_{i=m}^{n} \mu_{x_i}(\{b_i\}).$$

A channel ν is *deterministic* if there exists a mapping $f\colon A \to B$ such that

$$\nu_x(\{f(x)\}) = 1 \qquad \text{all } x.$$

Such a channel is stationary if and only if $f(Tx) = Tf(x)$, for all x.

A channel ν is *primitive* (Neuhoff and Shields 1979) if there exists an I.I.D. process Z, called the noise source, and a function f, called the sliding-block encoder, such that the output at time n is given by the formula

$$Y_n = f(T^n Z, T^n X),$$

and for some finite, positive integer L, called the coding half-width,

$$f(z, x) = f(\bar{z}, \bar{x}) \qquad \text{whenever} \quad z_{-L}^{L} = \bar{z}_{-L}^{L}, \quad x_{-L}^{L} = \bar{x}_{-L}^{L}.$$

Such a channel can be thought of as the cascade of the memoryless channel that outputs (Z, X) followed by the deterministic channel defined by $Y_n = f(Z_{n-L}^{n+L}, X_{n-L}^{n+L})$.

A stationary channel ν is *$\bar{d}$-continuous* (Gray and Ornstein 1979) if for any $\varepsilon > 0$ there is an N_0 such that

$$\bar{d}_n(\nu_x, \nu_{\bar{x}}) < \varepsilon \qquad \text{whenever} \quad n \geqslant N_0 \quad \text{and} \quad x^n = \bar{x}^n.$$

Such channels have a decaying input memory. A stationary channel is CABI (Neuhoff and Shields 1979) if for any $\varepsilon > 0$ there is an N_0 and for any $n \geqslant N_0$ there is an M_0 such that for any x and $m \geqslant M_0$

$$\bar{d}_m(\nu_x, [\nu_x]^n) \leqslant \varepsilon,$$

where $[\nu_x]^n$ denotes the product of the measures $\dots, (\nu_x)^0_{-n+1}, (\nu_x)^n_1, (\nu_x)^{2n}_{n+1}, \dots$. The measure $[\nu_x]^n$ is called the independent n-blocking of ν_x.

III. Channel Distances

In this section we define several channel distance measures and state our result concerning their properties and relationships. The deepest of these results (stated as (8b), (8c)) shows that for $\bar{d}$-continuous channels the strongest and weakest distances are uniformly equivalent but not identical. All these results are established in Section VI, except for the final result, Property (11), which is a consequence of a representation theorem proved in Section V. Let $\nu = [A, B, \nu, X, Y]$ and $\hat{\nu} = [A, B, \hat{\nu}, \hat{X}, \hat{Y}]$ be stationary channels with identical alphabets. We define the following concepts of channel distance:

(CD1) $\bar{D}(\nu, \hat{\nu}) = \limsup_{n\to\infty} \sup_{x} \bar{d}_n(\nu_x, \hat{\nu}_x)$,

(CD2) $\underline{D}(\nu, \hat{\nu}) = \sup_{x} \limsup_{n\to\infty} \bar{d}_n(\nu_x, \hat{\nu}_x)$,

(CD3) $D_S(\nu, \hat{\nu}) = \sup_{\alpha \text{ stationary}} \lim_{n\to\infty} \int \bar{d}_n(\nu_x, \hat{\nu}_x)\, d\alpha(x)$,

(CD4) $D_B(\nu, \hat{\nu}) = \sup_{\alpha \text{ block stationary}} \lim_{n\to\infty} \bar{d}_n(\alpha\nu, \alpha\hat{\nu})$,

(CD5) $D(\nu, \hat{\nu}) = \sup_{\alpha \text{ stationary}} \lim_{n\to\infty} \bar{d}_n(\alpha\nu, \alpha\hat{\nu})$.

The first distance $\bar{D}$ was used in (Neuhoff and Shields 1979, 1982a, and 1982b), where it was denoted by $\bar{d}$. The second distance $\underline{D}$ is obtained by switching the supremum and limit superior. In the third, D_{S}, and fifth, D, the supremum is over the class of stationary sources, while in the fourth, D_{B}, the supremum is over the class of block-stationary sources.

We now state a number of properties. With the exception of the last, they are proved in Section VI.

(1) In the definition of D_{S} the limit exists and can be interchanged with the supremum over the class of stationary sources. With either ordering, the limit can be replaced by a supremum over n, and the supremum over the class of stationary sources can be replaced by a supremum over the class of stationary ergodic sources or over the class of block-stationary sources.

(2) In the definition of D_{B}, the limit exists.

(3) In the definition of D the limit exists and can be interchanged with the supremum over the class of stationary sources. With either ordering, the limit can be replaced by a supremum over n, and the supremum over the class of stationary sources can be replaced by a supremum over the class of stationary ergodic sources.

(4) Each distance is a pseudometric on the class of all stationary channels, but not a metric.

(5) $\bar{D} \geqslant \underline{D} \geqslant D_{\mathrm{S}} \geqslant D_{\mathrm{B}} \geqslant D$.

(6) With the exception of D_{S} and D_{B}, any pair of the above distances are nonequivalent. (Two pseudometrics are equivalent if convergence in one implies convergence in the other.) We do not know whether D_{S} and D_{B} are identical or equivalent.

(7) On the class of memoryless channels:

(a) all the distances equal $\bar{D}$.

(b) $\bar{D}$ is a metric.

(8) On the class of $\bar{d}$-continuous channels:

(a) $\bar{D} = \underline{D} = D_{\mathrm{S}} = D_{\mathrm{B}}$.

(b) $\bar{D}$ is uniformly equivalent to D; that is, there is a constant c such that $\bar{D} \leqslant cD$.

(c) There exist ν and $\hat{\nu}$ such that $\bar{D} > D$.

(d) None of the distances are metrics.

(9) Let $(\mathbf{C}, D_{\mathrm{B}})$ denote the pseudometric space of all stationary channels with pseudometric D_{B}, let $(\mathbf{P}, \bar{d})$ denote the metric space of all pair processes on $A \times B$ and for any source α let ψ_α denote the mapping $\nu \to \alpha\nu$ from $(\mathbf{C}, D_{\mathrm{B}})$ into $(\mathbf{P}, \bar{d})$. Then

(a) The family of mappings $\{\psi_\alpha\}$ is equicontinuous on the class of block-stationary sources.

(b) The result (a) also holds if D_{B} is replaced by any of the stronger distances but not if replaced by D.

(c) The family of mapping $\{\psi_\alpha\}$ from $(\mathbf{C}, D)$ into $(\mathbf{P}, \bar{d})$ is equicontinuous on the class of stationary sources.

(10) If ν and $\hat{\nu}$ are stationary channels, then the following are equivalent statements:

(a) ν and $\hat{\nu}$ are equivalent; that is, $\bar{d}(\alpha\nu, \alpha\hat{\nu}) = 0$, for any block-stationary source α.
(b) $D_{\mathrm{B}} = 0$.
(c) $D = 0$.
(d) $D_{\mathrm{S}} = 0$.
(e) For any stationary source α, $\bar{d}(\alpha\nu, \alpha\hat{\nu}) = 0$.

(11) The class of $\bar{d}$-continuous, CABI channels is complete with respect to $\bar{D}$ and, consequently, with respect to $\underline{D}$, D_{S}, D_{B}, and D as they are all either equal or uniformly equivalent to $\bar{D}$ on this class.

IV. Approximation

As mentioned earlier, the various channel distances are intended to measure the degree to which one channel approximates another. An ideal distance concept would be strong enough that close channels have similar behavior but weak enough that channels without significant differences are lumped together. From Section III Property (9) we see that all the distances except the D-distance have the property that the family of mappings $\{\psi_\alpha\}$, $\psi_\alpha = \alpha\nu$, is equicontinuous on the class of block-stationary sources α. This guarantees that close channels will have similar capacities and similar performances when any block, convolutional, or sliding-block code is applied. While the D-distance has the property that $\{\psi_\alpha\}$ is equicontinuous on the class of stationary sources, this is not sufficient to guarantee that close channels have similar performances for block or convolutional codes. On the other hand Property (10) shows that only D, D_{B}, and D_{S} are weak enough that equivalent channels are assigned zero distance.

The above discussion suggests that D_{B} is the most suitable distance for measuring the degree to which one channel approximates another. Let us note, however, that for the important class of $\bar{D}$-continuous channels it does not matter which distance is chosen, for, as asserted in Property (8), they are all identical or at least uniformly equivalent. Finally, we note that the strongest distance $\bar{D}$ is the easiest to work with, for it is defined in terms of input sequences rather than sources.

We now turn to the question of what channels can be arbitrarily well approximated by primitive channels relative to the various distances. For the $\bar{D}$-distance it is the class of $\bar{d}$-continuous, CABI channels (Neuhoff and Shields 1979). For any of the weaker distance it is simply the closure of the $\bar{d}$-continuous, CABI channels. Since Property (11) shows that the $\bar{d}$-continuous, CABI channels are complete relative to any of the distances, their closure is obtained simply by adding all the channels at distance zero. Hence we have

THEOREM 4.1. *For the $\underline{D}$-, D_S-, D_B-, or D-distance, the class of channels that can be arbitrarily well approximated by primitive channels relative to the given distance equals the class of $\bar{d}$-continuous channels plus all other channels at distance zero from some $\bar{d}$-continuous,* CABI *channel.*

The examples used to prove Property (6) of Section III also show that for the $\underline{D}$-distance the closure class is larger than the class of $\bar{d}$-continuous, CABI channels, and for the D_S-distance the closure class is larger still. Property (10) of Section III shows that the closure class is that same for D_S, D_B, and D and equals the $\bar{d}$-continuous, CABI channels plus all equivalent channels.

V. COMPLETENESS AND EXACT REPRESENTATION

In this section we show that the class of $\bar{d}$-continuous, CABI channels is $\bar{D}$-complete. Our proof also shows that the ouput of such a channel can be represented as an infinite length sliding-block coding of the input and an independent noise source. The key to these results is

LEMMA 5.1. *Let μ be a $\bar{d}$-continuous,* CABI *channel and let ν be a primitive channel with noise source Z and sliding-block encoder f such that $\bar{D}(\mu, \nu) < \varepsilon$. Given $\delta > 0$ and a nontrivial binary* I.I.D. *process R, independent of Z, there is a primitive channel $\hat{\nu}$ with noise source (Z, R) and sliding-block encoder $\hat{f}$ such that*

(i) $\bar{D}(\mu, \hat{\nu}) < \delta$,

(ii) $\operatorname{Prob}(f(Z, x) \neq \hat{f}(Z, R, x)) < \varepsilon + \delta$, *all* $x \in A$.

Proof. We first show that given $\alpha > 0$ there is an $N_1 = N_1(\alpha)$ such that if $n \geqslant N_1$ and $a_1^n \in A_1^n$ there is a function $Y_1^n = \varphi_n(z_1^n, a_1^n)$ with distribution $\bar{\mu}_{a_1^n}$ such that if $\bar{x}_1^n = a_1^n$, then

(a) $\bar{d}_n(\bar{\mu}_{a_1^n}, \mu_{\bar{x}}^n) < \alpha$,

(b) $1/n \sum_{i=1}^n \operatorname{Prob}(Y_i \neq f(T^i Z, T^i \bar{x})) < \varepsilon + \alpha$.

The integer N_1 is chosen so large that

(1) $2l/N_1 < \alpha$,

where l is the coding half-width for the sliding-block encoder f, and so that if $n \geqslant N_1$, then the following two conditions hold:

(2) $\bar{d}_n(\nu_x, \mu_x) < \varepsilon$ for all x,

(3) $\bar{d}_n(\mu_x, \mu_{\bar{x}}) < \alpha$ if $x_1^n = \bar{x}_1^n$.

Condition (2) uses the assumption that $\bar{D}(\nu, \mu) < \varepsilon$, while condition (3) uses the assumption that μ is $\bar{d}$-continuous.

To construct the function φ_n we fix a_1^n, choose x such that $x_1^n = a_1^n$ and let U_1 denote the unit interval $[0, 1]$. For each integer i in the range $l+1 \leqslant i \leqslant n-l$ we define the partition $Q^{(i)} = \{Q_b^{(i)}\colon b \in B_1\}$ of the n-dimensional cube U_1^n by

(4) $Q_b^{(i)} = \{z_1^n : f(T^i \bar{z}, T^i x) = b, \text{ whenever } \bar{z}_1^n = z_1^n\}$

For those integers i in the range $1 \leqslant i \leqslant l$ and $n-l < i \leqslant n$ the partitions $Q^{(i)} = \{Q_b^{(i)}\colon b \in B_1\}$ are defined arbitrarily. Let $\bar{\nu}$ be the measure on B_1^n defined by

$$\bar{\nu}(b_1^n) = \lambda\left(\bigcap_{i=1}^{n} Q_{b_i}^{(i)}\right),$$

where λ denotes Lebesgue measure on U_1^n. Condition (1) guarantees that

(5) $\bar{d}_n(\bar{\nu}, \nu_x) < \alpha$.

Next we use the partition definition of $\bar{d}_n$ (Property (A.1)) together with conditions (2) and (5) to choose partitions $P^{(i)} = \{P_b^{(i)}\colon b \in B_1\}$ of U_1^n such that the following two conditions hold:

(6) $\lambda(\bigcap_{i=1}^n P_{b_i}^{(i)}) = \mu_x(b_1^n)$,

(7) $(1/n) \sum_{i=1}^n |P^{(i)} - Q^{(i)}|_\lambda < \varepsilon + \alpha$,

where the notation used in (7) is defined in (A.1). We then define

$$\varphi_n(z_1^n, a_1^n) = b_1^n \qquad \text{if} \quad z_1^n \in \bigcap_{i=1}^{n} P_{b_i}^{(i)}.$$

Condition (6) guarantees that the distribution $\bar{\mu}_{a_1^n}$ of φ_n is the same as μ_x and hence condition (3) guarantees that the distribution property (a) holds. The definition (4) of $Q^{(i)}$ together with condition (7) guarantees that property (b) holds.

The function φ_n defines a block code of length n from $U_1^n \times A_1^n$ into B_1^n. This block code can be used to define a sliding-block code which in turn defines the desired approximating channel $\bar{\nu}$ as shown in our earlier paper (Neuhoff and Shields 1979, Appendix C). We sketch the idea here, referring the reader to our earlier paper for details.

Let us fix $N \geqslant N_1$ and choose a cylinder set E in the R process of such low probability that the waiting time τ between occurrences of E is, with high probability, very large relative to N. We then fix the sequence x and apply the block code φ_N to successive blocks of length N from x following the occurrence of E in Z. That is if n_1 and n_2 are successive occurrences of E we define

$$\begin{aligned} b_{n_1}^{n_1+N-1} &= \varphi_N(Z_{n_1}^{n_1+N-1}, x_{n_1}^{n_1+N-1}), \\ b_{n_1+N}^{n_1+2N-1} &= \varphi_N(Z_{n_1+N}^{n_1+2N-1}, x_{n_1+N}^{n_1+2N-1}), \\ &\vdots \\ b_{n_1+(k-1)N}^{n_2+kN-1} &= \varphi_N(Z_{n_1+(k-1)N}^{n_1+kN-1}), x_{n_1+(k-1)N}^{n_1+kN-1}), \end{aligned}$$

where k is the largest integer less than some fixed integer K (to be specified later) such that

$$n_1 + kN - 1 \leqslant n_2,$$

and we define

$$b_n = b' \qquad \text{if} \quad n_1 + kN \leqslant n \leqslant n_2,$$

where b' is some fixed letter. This defines a sliding-block code $\hat{f}$ from $U \times A$ into B which yields the desired primitive channel $\hat{\nu}$.

If the waiting time τ is sufficiently large and if the cut-off rule K is sufficiently large, then most of the output consists of blocks of length N that are conditionally independent, given x. Thus the CABI property guarantees that for suitable choice of α and sufficiently large N, the resulting channel $\hat{\nu}$ will be within δ of μ while property (b) guarantees that

$$\text{Prob}(f(Z, x) \neq \hat{f}(Z, R, x)) < \varepsilon + \delta, \qquad \text{all } x.$$

This proves the lemma.

We now use this lemma to establish $\bar{D}$-completeness (Theorem 5.3) and obtain a representation theorem (Theorem 5.4) for the class of $\bar{d}$-continuous, CABI channels. These two results are simple consequences of

LEMMA 5.2. *Suppose $\{\nu^{(n)}\}$ is a sequence of $\bar{d}$-continuous,* CABI *channels such that $\bar{D}(\nu^{(n)}, \nu^{(n+1)}) = \varepsilon_n$, where $\sum \varepsilon_n$ converges. Let Z be an* I.I.D. *noise source uniformly distributed on $U_1 = [0, 1]$. There is a measurable function f: $U \times A \to B$ such that if ν is the channel defined by*

$$Y_n = f(T^n Z, T^n X),$$

then

$$\lim_{n \to \infty} \bar{D}(\nu, \nu^{(n)}) = 0.$$

Proof. Let $\{\delta_n\}$ be a summable sequence of positive numbers. Also let V, $R^{(1)}$, $R^{(2)}$,..., be a sequence of I.I.D. processes, independent of each other, such that V_n is uniformly distributed on $[0, 1]$, each $R^{(i)}$ is binary and nontrivial, and V_n and each $R_n^{(i)}$ is a function of Z_n. We then define the I.I.D. processes $W^{(k)} = \{W_n^{(k)}\}$, $k = 1, 2,...$. according to

$$W_n^{(k)} = (Z_n, R_n^{(1)}, R_n^{(2)},..., R_n^{(k)}).$$

We can now use the approximation theorem of (Neuhoff and Shields 1979) to choose a primitive channel $\hat{\nu}^{(1)}$ with noise source $W^{(1)}$ and encoder f_1 such that

$$\bar{D}(\nu^{(1)}, \hat{\nu}^{(1)}) < \delta_1.$$

Since $\bar{D}(\nu^{(1)}, \nu^{(2)}) = \varepsilon_1$, we have

$$\bar{D}(\hat{\nu}^{(1)}, \nu^{(2)}) < \delta_1 + \varepsilon_1,$$

so we can apply Lemma 5.1 to obtain a primitive channel $\hat{\nu}^{(2)}$ with noise source $W^{(2)}$ and encoder f_2 such that

$$\bar{D}(\nu^{(2)}, \hat{\nu}^{(2)}) < \delta_2 \qquad \text{and} \qquad \text{Prob}(f_1 \neq f_2) < \delta_1 + \delta_2 + \varepsilon_1.$$

We can then proceed by induction to obtain primitive channels $\hat{\nu}^{(3)}, \hat{\nu}^{(4)},...,$ with respective noise sources $W^{(3)}$, $W^{(4)}$,..., and encoders $f_3, f_4,...,$ such that

(8) $\bar{D}(\nu^{(n+1)}, \hat{\nu}^{(n+1)}) < \delta_{n+1}$

and

(9) $\text{Prob}(f_{n+1}(Z, x) \neq f_n(Z, x)) < \delta_{n+1} + \delta_n + \varepsilon_n$, all x.

It follows from (9) that for each sequence x the limit

$$f(Z, x) = \lim_{n \to \infty} f_n(Z, x)$$

exists with probability 1 and from (8) that the channel ν defined by $Y_n = f(T^n Z, T^n X)$ is the $\bar{D}$-limit of $\hat{\nu}^{(n)}$. The triangle inequality

$$\bar{D}(\nu, \nu^{(n)}) \leqslant \bar{D}(\nu, \hat{\nu}^{(n)}) + \bar{D}(\hat{\nu}^{(n)}, \nu^{(n)})$$

shows that $\lim_n \bar{D}(\nu, \nu^{(n)}) = 0$, which proves Lemma 5.2.

THEOREM 5.3. *The class of $\bar{d}$-continuous*, CABI *channels is $\bar{D}$-complete.*

Proof. If $\{\nu^{(n)}\}$ is $\bar{D}$-Cauchy then we can drop to a subsequence, if necessary, to obtain

$$\bar{D}(\nu^{(n)}, \nu^{(n+1)}) < 2^{-n}.$$

If each $\nu^{(n)}$ is $\bar{d}$-continuous and CABI, we can use Lemma 5.2 to obtain the limit channel ν, which is necessarily $\bar{d}$-continuous and CABI (Neuhoff and Shields 1979).

THEOREM 5.4. *If ν is $\bar{d}$-continuous and* CABI *and Z is an* I.I.D. *noise source, uniformly distributed on* $[0, 1]$, *there is a measurable function $f(z, x)$ such that if $\hat{\nu}$ is the channel defined by*

$$\hat{Y}_n = f(T^n Z, T^n X),$$

then

$$\bar{D}(\nu, \hat{\nu}) = 0.$$

Proof. We just apply Lemma 5.2 with each $\nu^{(n)}$ equal to ν.

The converse of Theorem 5.4 is false; that is, there is a measurable function $f(z, x)$ such that if $\hat{\nu}$ is defined by $\hat{Y}_n = f(T^n Z, T^n X)$, then $\hat{\nu}$ is not $\bar{d}$-continuous, as shown in the Appendix of (Neuhoff and Shields 1979). One might hope that such "infinitely" primitive channels are $\bar{D}$-distance zero from the class of $\bar{d}$-continuous, CABI channels, which is the $\bar{D}$-closure of the primitive channels. This is not true, as pointed out to us by J.-P. Thouvenot, because of a result in Bailey (1976). A sequence x is typical of the stationary process α if each finite sequence a_1^l occurs in x with limiting relative frequency equal to $\alpha(a_1^l)$. One can show that the set E of sequences that are typical of some ergodic measure of entropy zero is a Borel set so that its indicator function $f = \Psi_E$ is measurable. Bailey showed that there is no sequence $\{f_n\}$ such that $f_n(x)$ depends only on x_1^n for which $f_n(x)$ converges almost everywhere to $f(x)$ for each ergodic measure α. Thus, if $\hat{\nu}$ is the channel defined by $\hat{Y}_n = f(T^n X)$, then $\bar{D}(\nu, \hat{\nu})$ must be positive for any $\bar{d}$-continuous, CABI channel ν.

VI. PROOFS FOR SECTION III

(1) If we let $\mathbf{S}$ denote the class of stationary sources α and let

$$a_n = \int \bar{d}_n(\nu_x, \hat{\nu}_x)\, d\alpha(x),$$

then the definition of D_{S} takes the form

$$D_{\mathrm{S}} = \sup_{\mathbf{S}} \lim_{n \to \infty} a_n.$$

If α, ν, and $\hat{\nu}$ are stationary, then Property (A.3a) can be used to show that

$\{a_n\}$ is superadditive; i.e., $na_n \geqslant ma_m + la_l$, whenever $n = m + l$. It follows (cf. Gallager 1968) that $\lim_n a_n$ exists and equals $\sup_n a_n$, and hence we can write

$$\begin{aligned} D_{\mathrm{S}} &= \sup_{\mathbf{S}} \sup_{n} \int \bar{d}_n(\nu_x, \hat{\nu}_x)\, d\alpha(x) \\ &= \sup_{n} \sup_{\mathbf{S}} \int \bar{d}_n(\nu_x, \hat{\nu}_x)\, d\alpha(x). \end{aligned}$$

Next we show that $\sup_n$ can be replaced by $\lim_n$ to obtain the formula

$$D_{\mathrm{S}}(\nu, \hat{\nu}) = \lim_{n} \sup_{\mathbf{S}} \int \bar{d}_n(\nu_x, \hat{\nu}_x)\, d\alpha(x). \tag{6.1}$$

We establish this by using the superadditivity of $\{a_n\}$ to obtain

$$na_n \geqslant kma_{km} + la_l \geqslant kma_m,$$

if $n = km + l$. If $l < m$, then we can write

$$na_n \geqslant kma_m \geqslant na_m - m.$$

Thus if $b_n = \sup_{\mathbf{S}} a_n$ and $n \geqslant m$, then we have

$$b_n \geqslant b_m - \frac{m}{n},$$

so that $\sup_n b_n = \lim_n b_n$, which proves (6.1).

We will now show that replacing $\mathbf{S}$ by the larger class $\mathbf{B}$ of block stationary sources does not increase D_{S}. For any N let α be an N-stationary source and let β be the stationary source defined by the formula

$$\beta(E) = \sum_{i=0}^{N-1} \frac{1}{N} \alpha(T^i E).$$

Then for any n

$$\begin{aligned} \int \bar{d}_n(\nu_x, \hat{\nu}_x)\, d_{\beta}(x) &= \sum_{i=0}^{N-1} \frac{1}{N} \int \bar{d}_n(\nu_{T^{-i}x}, \hat{\nu}_{T^{-i}x}\, d\alpha(x) \\ &= \int \bar{d}_n(\nu_x, \hat{\nu}_x)\, d\alpha(x), \end{aligned}$$

where we have used the stationarity of ν and $\hat{\nu}$. It follows that the original definition, (CD3), can be written in the form

$$D_{\mathrm{S}}(\nu, \hat{\nu}) = \sup_{\alpha \in \mathbf{B}} \lim_{n} \int \bar{d}_n(\nu_x, \hat{\nu}_x)\, d\alpha(x). \tag{6.2}$$

We now show that if **S** is replaced in (CD3) by the smaller class **E** of ergodic sources, then D_{S} does not decrease. If α is a nonergodic stationary source, then α has an ergodic decomposition $[\alpha_\theta, w]$. Hence

$$\int \bar{d}_n(\nu_x, \hat{\nu}_x)\, d\alpha(x) = \int \left\{ \int \bar{d}_n(\nu_x, \hat{\nu}_x)\, d\alpha_\theta(x) \right\} dw(\theta).$$

Since there must exist some θ such that the quantity in brackets is at least as large as the left-hand side of the above, it follows that D_{S} is not made smaller by replacing **S** by **E** in (CD3).

(2) For any block-stationary α, $\lim_n \bar{d}_n(\alpha\nu, \alpha\hat{\nu})$ is just $\bar{d}(\alpha\nu, \alpha\hat{\nu})$. The limit cannot, however, be replaced by $\sup_n$.

(3) In the definition of $\bar{D}$, for any $\alpha \in \mathbf{S}$ $\lim_n \bar{d}_n(\alpha\nu, \alpha\hat{\nu})$ is just $\bar{d}(\alpha\nu, \alpha\hat{\nu})$, which by Property (A.2)(a), equals $\sup_n \bar{d}_n(\alpha\nu, \alpha\hat{\nu})$. Using the same technique as used in (1) for D_{S} one can show

$$\begin{aligned} D(\nu, \hat{\nu}) &= \sup_{n} \sup_{\alpha \in \mathbf{S}} \bar{d}_n(\alpha\nu, \alpha\hat{\nu}) \\ &= \lim_{n} \sup_{\alpha \in \mathbf{S}} \bar{d}_n(\alpha\nu, \alpha\hat{\nu}). \end{aligned}$$

We now show that if the class **S** of stationary sources is replaced by the smaller class **E** of ergodic sources, then D does not get smaller. Let α be a nonergodic source in **S** and let $[\alpha_\theta, w]$ be its ergodic decomposition. Then

$$\alpha\nu(E \times F) = \int \left[\int_E \nu_x(F)\, d\alpha_\theta(x) \right] dw(\theta),$$

and there is a similar expression for $\alpha\hat{\nu}$. Hence by the convexity of $\bar{d}_n$ (Property (A.4)(b))

$$\bar{d}_n(\alpha\nu, \alpha\hat{\nu}) \leqslant \int \bar{d}_n(\alpha_\theta \nu, \alpha_\theta \hat{\nu})\, dw(\theta).$$

Since there must exist some θ such that $\bar{d}_n(\alpha\nu, \alpha\hat{\nu}) \leqslant \bar{d}_n(\alpha_\theta \nu, \alpha_\theta \hat{\nu})$, it follows that D is not made smaller by replacing **S** by **E**.

(4) Each distance is obviously non-negative and symmetric. The triangle inequality for each follows directly from the triangle inequality for

$\bar{d}_n$. Hence, each is a pseudometric. An example of a pair of channels ν, $\hat{\nu}$ for which $\bar{D}(\nu, \hat{\nu}) = 0$ was given in (Neuhoff and Shields 1979, Appendix A). Since Property (5) shows that the other distances are smaller, they too are zero. Hence, none of them are metrics.

(5) The inequality $\bar{D} \geqslant \underline{D}$ is elementary. The inequality $\underline{D} \geqslant D_S$ follows deom Fatou's lemma. The inequality $D_S \geqslant D_B$ follows from Property (A.4)(a) of the Appendix. Finally the inequality $D_B \geqslant D$ is obvious.

(6) We demonstrate the nonequivalence of the various distances through a series of examples consisting entirely of binary deterministic channels with elphabets $A_1 = B_1 = \{0, 1\}$. Let us observe that if ν and $\hat{\nu}$ are deterministic channels with inputs x and $\hat{x}$ that produce outputs y and $\hat{y}$, respectively, then $\bar{d}_n(\nu_x, \hat{\nu}_x) = d_n(y, \hat{y})$. The binary *complement* a^c is defined by $a^c = 0$ iff $a = 1$ and we use $(u_m^n)^c$ to denote the sequence u_i^c, $m \leqslant i \leqslant n$.

(a) To show that $\bar{D}$ and $\underline{D}$ are not equivalent, we exhibit a pair of channels ν and $\hat{\nu}$ such that $\bar{D}(\nu, \hat{\nu}) = 1$ and $\underline{D}(\nu, \hat{\nu}) = 0$. Let ν be the identity channel; that is, the output Y equals the input X. Let $u^{(1)}, u^{(2)},\ldots,$ be a collection of aperiodic sequences in A such that no one is a shift of another (i.e., $T^i u^{(j)} \neq u^{(k)}$, all i, j, k with $j \neq k$). Let $v^{(j)}$ be the sequence such that $v_n^{(j)} = (u_n^{(j)})^c$ if $1 \leqslant n \leqslant j$, and $v_n^{(j)} = u_n^{(j)}$ if otherwise. Let $\hat{\nu}$ be the deterministic channel that produces $T^i v^{(j)}$ if $X = T^i u^{(j)}$ for some i, j and produces X if otherwise.

For any n there is an input sequence, namely $u^{(n)}$, for which the channels produce unequal outputs. In particular,

$$\bar{d}_n(\nu_{u^{(n)}}, \hat{\nu}_{u^{(n)}}) = d_n(u^{(n)}, v^{(n)}) = 1.$$

It follows that $\bar{D}(\nu, \hat{\nu}) = 1$. On the other hand for any input sequence x, the outputs are eventually identical. In particular, if $x = T^i u^{(n)}$, then

$$\bar{d}_m(\nu_x, \hat{\nu}_x) = d_m(u^{(n)}, v^{(n)}) \leqslant \frac{n}{m},$$

while if $x \neq T^i u^{(n)}$, all i, n, then $\bar{d}_m(\nu_x, \hat{\nu}_x) = 0$. Hence for any x

$$\limsup_{m \to \infty} \bar{d}_m(\nu_x, \hat{\nu}_x) = 0$$

and it follows that $\underline{D}(\nu, \hat{\nu}) = 0$.

(b) To show that $\underline{D}$ is not equivalent to D_S we exhibit a pair of channels ν and $\hat{\nu}$ such that $\underline{D} = 1$ and $D_S = 0$. Let ν be the identity channel. Let u be some aperiodic input sequence, let

$v = u^c$, and let $\hat{\nu}$ be the channel that produces $T^i v$ when $X = T^i u$ for some i and produces X otherwise. For any n

$$\bar{d}_n(\nu_u, \hat{\nu}_u) = d_n(u, v) = 1.$$

Hence $\underline{D}(\nu, \hat{\nu}) = 1$. On the other hand, for any $x \neq T^i u$, all i,

$$\bar{d}_n(\nu_x, \hat{\nu}_x) = 0 \qquad \text{all } n.$$

Now for any stationary α, the set $\{..., T^{-1}u, u, Tu,...\}$ has zero probability; hence

$$\int \bar{d}_n(\nu_x, \hat{\nu}_x)\, d\alpha(x) = 0.$$

Therefore $D_S(\nu, \hat{\nu}) = 0$.

(c) Let us note that by Property (10) $D_B = 0$ whenever $D = 0$. Hence it is not possible to demonstrate the nonequivalence of D_B and D as in parts (a) and (b). To show that D_B is not equivalent to D we exhibit a sequence of channels $\{\nu^{(k)}\}$ that converges to a channel ν in the D-distance, but not in the D_B-distance.

For any integer $L \geqslant 2$ let $u^{(L)}$ be the periodic sequence having period L and $(u_1^{(L)},..., u_L^{(L)}) = (0\ 1\ 1\ 1\ 1...)$. Let $v^{(L)}$ be the periodic sequence having period L and $(v_1^{(L)},..., v_L^{(L)}) = (0\ 1\ 0\ 1\ 0...)$. Let ν be the deterministic channel that produces $T^i v^{(L)}$ when $X = T^i u^{(L)}$ for some i and L and produces X otherwise. For any K let $\nu^{(K)}$ be the deterministic channel that produces $T^i v^{(L)}$ when $X = T^i u^{(L)}$ for some $L < K$ and some i and produces $w^{(L)} = (v^{(L)})^c$ when $X = T^i u^{(L)}$ for some $L \geqslant K$ and some i and produces X otherwise. Observe that the output of ν equals the output of $\nu^{(K)}$ except if X is a shift of $u^{(L)}$ for some $L \geqslant K$, in which case the outputs disagree in every place. Therefore for any n

$$\begin{aligned} \bar{d}(\nu_x, \nu_x^{(K)}) &= 1, \qquad x = T^i u^{(L)} \text{ some } i \text{ and } L \geqslant K, \\ &= 0, \qquad \text{otherwise.} \end{aligned} \tag{6.3}$$

We now show that $D_B(\nu, \nu^{(K)}) \to 1$ as $K \to \infty$. Fix $K \geqslant 3$ and let $\alpha^{(K)}$ be the K-stationary source that assigns probability one to $u^{(K)}$. Then for any n Eq. (6.3) implies

$$\bar{d}_n(\alpha^{(K)}\nu, \alpha^{(K)}\nu^{(K)}) = \bar{d}_n(\nu_{u^{(K)}}, \nu_{u^{(K)}}^{(K)}) = 1.$$

Hence $D_B(\nu, \nu^{(K)}) = 1$ for all $K \geqslant 3$.

To show that $D(\nu, \nu^{(K)}) \to 0$ as $K \to \infty$, we will show that for any K and any stationary source α,

$$\bar{d}(\alpha\nu, \alpha\nu^{(K)}) \leqslant \frac{3}{K}.$$

Note that Property (3) implies that it is sufficient to consider only ergodic α. So let us fix K, n, and α and let $\hat{\nu} = \nu^{(K)}$. Since α is ergodic, the invariant sets $E^{(L)} = \{T^i u^{(L)}: 0 \leqslant i \leqslant L-1\}$ either all have probability zero or exactly one has probability one. In the former case Property (A.4)(a) and (6.3) imply that for any n

$$\bar{d}_n(\alpha\nu, \alpha\hat{\nu}) \leqslant \int \bar{d}_n(\nu_x, \hat{\nu}_x)\, d\alpha(x) = 0. \tag{6.4}$$

Hence $\bar{d}(\alpha\nu, \alpha\hat{\nu}) = 0$. In the latter case there is exactly one L such that $\alpha(\{T^i u^{(L)}\}) = 1/L$, $i = 0, \ldots, L-1$. If $L < K$, then just as in (6.4) $\bar{d}(\alpha\nu, \alpha\hat{\nu}) = 0$. On the other hand if $L \geqslant K$, we can show that $\bar{d}(\alpha\nu, \alpha\hat{\nu})$ is small by joining $\alpha\nu$ and $\alpha\hat{\nu}$ so that the input X to ν equals $T\hat{X}$, where $\hat{X}$ is the input to $\hat{\nu}$. In particular let $\omega \in \alpha\nu \nabla \alpha\hat{\nu}$ be the stationary measure on $A \times B \times A \times B$ such that

$$\omega(\{T^i u^{(L)}\} \times \{T^i v^{(L)}\} \times \{T^{i+1} u^{(L)}\} \times \{T^{i+1} w^{(L)}\}) = \frac{1}{L}, \qquad i = 0, \ldots, L-1.$$

Then by Property (A.2)(b)

$$\begin{aligned} \bar{d}(\alpha\nu, \alpha\hat{\nu}) &\leqslant E_\omega d_L((XY), (\hat{X}\hat{Y})) \\ &\leqslant E_\omega d_L(X, \hat{X}) + E_\omega d_L(Y, \hat{Y}) \\ &= d_L(u^{(L)}, Tu^{(L)}) + d_L(v^{(L)}, Tw^{(L)}) \\ &\leqslant \frac{2}{L} + \frac{1}{L} \\ &= \frac{3}{L} \leqslant \frac{3}{K}. \end{aligned}$$

In conclusion we have shown that for any ergodic α, $\bar{d}(\alpha\nu, \alpha\nu^{(K)}) \leqslant 3/K$. It follows that $D(\nu, \nu^{(K)}) \leqslant 3/K$ and, finally, that $D(\nu, \nu^{(K)}) \to 0$ as $K \to \infty$.

(7a) Suppose ν and $\hat{\nu}$ are memoryless. Since $\bar{D} \geqslant \underline{D} \geqslant D_{\mathrm{S}} \geqslant D_{\mathrm{B}} \geqslant D$, it suffices to show that $\bar{D} \geqslant D$. Since ν and $\hat{\nu}$ are memoryless, for any $a \in A_1$, there are measures μ_a and $\hat{\mu}_a$ on B_1 such that for any x with $x_1 = a$, μ_a and $\hat{\mu}_a$ equal ν_x^1 and $\hat{\nu}_x^1$, respectively. Let us choose a to maximize $\bar{d}_1(\mu_a, \hat{\mu}_a)$ and let $u = (\ldots a, a, a, \ldots)$. Then for any x and n the memoryless property implies (see Property (A.3)(b))

$$\bar{d}_n(\nu_x, \hat{\nu}_x) = \frac{1}{n} \sum_{i=1}^{n} \bar{d}_1(\mu_{x_i}, \hat{\mu}_{x_i}) \leqslant \bar{d}_1(\mu_a, \hat{\mu}_a).$$

It follows that $\bar{D}(\nu, \hat{\nu}) \leqslant \bar{d}_1(\mu_a, \hat{\mu}_a)$. Now let α be the stationary source such that $\alpha(\{u\}) = 1$. Then

$$D(\nu, \hat{\nu}) \geqslant \bar{d}(\alpha\nu, \alpha\hat{\nu}) = \bar{d}(\nu_u, \hat{\nu}_u) = \bar{d}_1(\mu_a, \hat{\mu}_a) \geqslant \bar{D}(\nu, \hat{\nu}).$$

(7b) From part (a) we see that if $\bar{D}(\nu, \hat{\nu}) = 0$, then $\bar{d}_1(\mu_a, \hat{\mu}_a) = 0$. This means $\mu_a = \hat{\mu}_a$, which in turn implies that $\nu = \hat{\nu}$. Hence $\bar{D}$ is a metric for memoryless channels.

(8a) Suppose ν and $\hat{\nu}$ are $\bar{d}$-continuous. Since $\bar{D} \geqslant \underline{D} \geqslant D_{\mathrm{S}} \geqslant D_{\mathrm{B}}$ it suffices to show that $\bar{D} \leqslant D_{\mathrm{B}}$. Given $\varepsilon > 0$ we use the $\bar{d}$-continuity property of both ν and $\hat{\nu}$ to choose N so that if $x^N = \bar{x}^N$, then

$$\bar{d}_N(\nu_x, \nu_{\bar{x}}) \leqslant \frac{\varepsilon}{3} \qquad \text{and} \qquad \bar{d}_N(\hat{\nu}_x, \hat{\nu}_{\bar{x}}) \leqslant \frac{\varepsilon}{3}. \tag{6.5}$$

In addition we may choose N so there exists $u \in A$ such that

$$\bar{d}_N(\nu_u, \hat{\nu}_u) \geqslant \bar{D}(\nu, \hat{\nu}) - \frac{\varepsilon}{3}. \tag{6.6}$$

Let v be the periodic sequence with period N such that $v_{iN+1}^{iN+N} = u_1^N$ for all i. Then (6.5), (6.6), and the triangle inequality imply

$$\bar{d}_N(\nu_v, \hat{\nu}_v) \geqslant \bar{D}(\nu, \hat{\nu}) - \varepsilon. \tag{6.7}$$

Since v is periodic, ν_v and $\hat{\nu}_v$ are N-stationary and Property (A.2)(a) implies

$$\bar{d}(\nu_v, \hat{\nu}_v) \geqslant \bar{d}_N(\nu_v, \hat{\nu}_v). \tag{6.8}$$

Now let α be the N-stationary source such that $\alpha(\{v\}) = 1$. Then by the definition of D_{B}

$$D_{\mathrm{B}}(\nu, \hat{\nu}) \geqslant \bar{d}(\alpha\nu, \alpha\hat{\nu}) = \bar{d}(\nu_v, \hat{\nu}_v) \geqslant \bar{D}(\nu, \hat{\nu}) - \varepsilon,$$

where we have used (6.7) and (6.8). Since ε is arbitrary, we have

$$D_{\mathrm{B}}(\nu, \hat{\nu}) \geqslant \bar{D}(\nu, \hat{\nu}).$$

(8b) To show that D and $\bar{D}$ are uniformly equivalent we will show that if ν and $\hat{\nu}$ are $\bar{d}$-continuous, then

$$D(\nu, \hat{\nu}) \geqslant \frac{\bar{D}(\nu, \hat{\nu})}{18}. \tag{6.9}$$

Fix ε, $0 < \varepsilon < \bar{D}(\nu, \hat{\nu})/3$. By the $\bar{d}$-continuity of ν and $\hat{\nu}$ we can choose N_0 so that if $x^{N_0} = \bar{x}^{N_0}$, then

$$\bar{d}_{N_0}(\nu_x, \nu_{\bar{x}}) \leqslant \varepsilon \qquad \text{and} \qquad \bar{d}_{N_0}(\hat{\nu}_x, \hat{\nu}_x) \leqslant \varepsilon. \tag{6.10}$$

In addition let us choose N_0 to be 2^{j-1} for some positive integer j so large that there exists $u \in A$ such that

$$\bar{d}_{N_0}(\nu_u, \hat{\nu}_u) \geqslant \bar{D}(\nu, \hat{\nu}) - \varepsilon. \tag{6.11}$$

We now claim there exists a periodic sequence v with period $N = 3N_0$ such that two inequalities hold

$$\bar{d}(\nu_v, \hat{\nu}_v) \geqslant \frac{\bar{D}(\nu, \hat{\nu})}{3} - \varepsilon \tag{6.12}$$

$$d(v, T^k v) \geqslant \tfrac{1}{6}, \qquad 0 < k < N. \tag{6.13}$$

The idea is to choose $\underline{w} = (w_1 \cdots w_{N_0})$ that is far apart from cyclic shifts of itself and let $\underline{v} \triangleq (v, \ldots, v_N) = (\underline{u}, \underline{w}, \underline{w})$, where $\underline{u} \triangleq (u_1, \ldots, u_{N_0})$. (The cyclic shift of $\underline{w}$ is $(w_2, w_3, \ldots, w_{N_0}, w_1)$.) With $\underline{v}$ so chosen, Properties (A.2)(a) and (A.3)(a) imply

$$\bar{d}(\nu_v, \hat{\nu}_v) \geqslant \bar{d}_N(\nu_v, \hat{\nu}_v) \geqslant \tfrac{1}{3}\bar{d}_{N_0}(\nu_v, \hat{\nu}_v). \tag{6.14}$$

Equation (6.12) then follows directly from (6.10), (6.11), (6.14), and the triangle inequality. To demonstrate (6.13) we arbitrarily choose two letters a and b from the alphabet A_1 and let $\underline{w}$ be a maximal length binary shift register sequence with length N_0 and with a's and b's as components (cf. Gallager 1968, pp. 230, 231). It is well known that any such sequence differs from any cyclic shift of itself in $(N_0 + 1)/2$ places. Furthermore we may assume that $\underline{w}$ differs from $\underline{u}$ in at least $N/2$ places, for otherwise we could replace $\underline{w}$ by its complement (i.e., interchange a's and b's) and retain the property that it differs from cyclic shifts of itself in at least $(N_0 + 1)/2$ places. To prove (6.13) we first observe that the periodicity of v implies

$$d(v, T^k v) = d_N(\underline{v}, S^k \underline{v}),$$

where $S\underline{v}$ denotes the cyclic shift of $\underline{v}$. To compute $d_N(\underline{v}, S^k \underline{v})$ one must count the number of places where $\underline{v}$ and $S^k \underline{v}$ disagree and divide by N. If k is a multiple of N_0, then imbedded in this count is a count of the number of disagreements between $\underline{u}$ and $\underline{w}$, which is at least $N_0/2$. If k is not a multiple of $N_0/2$, then imbedded in this count is a count of the number of disagreements between $\underline{w}$ and some cyclic shift of $\underline{w}$. In any case the number of disagreements is at least $N_0/2$ and (6.13) follows.

Having proved (6.12) and (6.13), our goal now is to show

$$D(\nu, \hat{\nu}) \geqslant \frac{\bar{d}(\nu_v, \hat{\nu}_v)}{6}. \tag{6.15}$$

Since ε is arbitrary, (6.12) and (6.15) imply $D(\nu, \hat{\nu}) \geqslant \bar{D}(\nu, \hat{\nu}) \geqslant \bar{D}(\nu, \hat{\nu})/18$, which is the desired result.

To prove (6.15), consider the stationary source α such that $\alpha(\{T^i v\}) = 1/N$. We will show that for any stationary joining λ of $\alpha\nu$ and $\alpha\hat{\nu}$

$$E_\lambda d_N((XY), (\hat{X}\hat{Y})) \geqslant \frac{\bar{d}(\nu_v, \hat{\nu}_v)}{6}. \tag{6.16}$$

Since $D \geqslant \bar{d}(\alpha\nu, \alpha\hat{\nu})$, Eq. (6.15) follows from (6.16) and Property (A.2)(b). The idea behind (6.16) is that any joining must make X and $\hat{X}$ equal to shifts of v. If the linking makes $X = \hat{X}$ with significant probability, then (6.12) implies $E_\lambda d_N$ is large. On the other hand if X is a shift of $\hat{X}$ with significant probability, then (6.13) implies $E_\lambda d_N$ is large. We now give the details.

First consider the case where $\alpha\nu$ and $\alpha\hat{\nu}$ are ergodic. Then by Property (A.2)(c) it is sufficient to assume λ is ergodic. For $i = 0,\ldots, N-1$, let $E_i = \{T^i v\} \times B$. Notice that the sets $E_i \times E_i$ are disjoint, that $E_i \times E_i = (T \times T)^i$ $(E_0 \times E_0)$ and that their union is invariant under $T \times T$. Since λ is ergodic, their union has probability 0 or 1. In the former case $X \neq \hat{X}$ with probability one and so

$$E_\lambda d_N((XY), (\hat{X}\hat{Y}) \geqslant E_\lambda d_N(X^N, \hat{X}^N) \geqslant \tfrac{1}{6}, \tag{6.17}$$

where the last inequality follows from (6.13). In the latter case $X = \hat{X}$ with probability one and so

$$E_\lambda d_N((XY), (\hat{X}\hat{Y}) = E_\lambda d_N(Y, \hat{Y}) \tag{6.18}$$

Let ω_i be the measure on $B \times B$ induced by λ restricted to $E_i \times E_i$. That is,

$$\omega_i(F \times \hat{F}) = \frac{\lambda(\{T^i v\} \times F \times \{T^i v\} \times \hat{F})}{\lambda(E_i \times E_i)}, \qquad F, \hat{F} \in B. \tag{6.19}$$

Since the sets $\{E_i \times E_i\}_{i=0}^{N \times 1}$ are disjoint, are shifts of one another, and have total measure 1, $\lambda(E_i \times E_i) = N^{-1}$, all i. It is then easy to show that ω_i is an N-stationary joining of $\nu_{T^i v}$ and $\hat{\nu}_{T^i v}$, and it follows that

$$\begin{aligned} E_\lambda d_N(Y, \hat{Y}) &= \frac{1}{N} \sum_{i=0}^{N-1} E_{\omega_i} d_N(Y, \hat{Y}) \\ &\geqslant \frac{1}{N} \sum_{i=0}^{N-1} \bar{d}(\nu_{T^i v}, \hat{\nu}_{T^i v}) \\ &= \bar{d}(\nu_v, \hat{\nu}_v), \end{aligned} \tag{6.20}$$

where the last equality follows Property (A.5). Together (6.17), (6.18), and (6.20) imply

$$E_\lambda d_N((XY), (\hat{X}\hat{Y}) \geqslant \min \left\{ \frac{1}{6}, \bar{d}(\nu_v, \hat{\nu}_v) \right\} \geqslant \frac{\bar{d}(\nu_v, \hat{\nu}_v)}{6}, \tag{6.21}$$

which is the desired result (6.16).

Finally we prove (6.12) assuming $\alpha\nu$ and $\alpha\hat{\nu}$ are stationary but not necessarily ergodic. In this case $\alpha\nu$ and $\alpha\hat{\nu}$ have ergodic decompositions $[(\alpha\nu)_\theta, w]$ and $[(\alpha\hat{\nu})_\varphi, \hat{w}]$, and the ergodic decomposition theory of Rohlin (1962) also shows there is a measure $r \in w \nabla \hat{w}$ and a measurable mapping $(\theta, \varphi) \to \lambda_{\theta,\varphi} \in (\alpha\hat{\nu})_\theta \nabla (\alpha\hat{\nu})_\varphi$ such that $\lambda_{\theta,\varphi}$ is stationary and ergodic and for any $F \in \mathbf{A} \times \mathbf{B} \times \mathbf{A} \times \mathbf{B}$,

$$\lambda(F) = \int \lambda_{\theta,\varphi}(F)\, dr(\theta, \varphi).$$

Let G be the collection of all θ, φ such that the invariant set $\bigcup_i (E_i \times E_i)$ has $\lambda_{\theta,\varphi}$ measure 1. Then

$$E_\lambda[d_N] = \int_G E_{\lambda_{\theta,\varphi}}[d_N]\, dr(\theta, \varphi) + \int_{G^c} E_{\lambda_{\theta,\varphi}}[d_N]\, dr(\theta, \varphi), \tag{6.22}$$

where d_N denotes $d_N((XY), (\hat{X}\hat{Y}))$. The argument leading to (6.17) shows that the second integral above can be bounded as

$$\int_{G^c} E_{\lambda_{\theta,\varphi}}[d_N]\, dr(\theta, \varphi) \geqslant \frac{r(G^c)}{6}. \tag{6.23}$$

We now consider the first integral in (6.22). For $(\theta, \varphi) \in G$, let $\omega_{\theta,\varphi,i}$ be defined as in (6.19), but with $\lambda_{\theta,\varphi}$ replacing λ. It is straightforward to show that $\omega_{\theta,\varphi,i}$ is an N-stationary joining of the N-stationary processes $\nu^\theta_{T^i v}$ and $\hat{\nu}^\varphi_{T^i v}$, where

$$\nu^\theta_{T^i v}(F) \triangleq \frac{(\alpha\nu)_\theta\, (\{T^i v\} \times F)}{\alpha(\{T^i v\})} = N(\alpha\nu)_\theta\, (\{T^i v\} \times F)$$

and where $\hat{\nu}^{\varphi}_{T^i v}$ is defined similarly. The argument leading to (6.20) shows that for $(\theta, \varphi) \in G$,

$$E_{\lambda_{\theta,\varphi}}[d_N] \geqslant \bar{d}(\nu^{\theta}_{v}, \hat{\nu}^{\varphi}_{v}).$$

Substituting this relation into the first integral in (6.22) gives

$$\int_G E_{\lambda_{\theta,\varphi}}[d_N]\, dr(\theta, \varphi) \geqslant \int \bar{d}(\nu^{\theta}_{v}, \hat{\nu}^{\varphi}_{v})\, dr(\theta, \varphi) - \int_{G^{c}} \bar{d}(\nu^{\theta}_{v}, \hat{\nu}^{\theta}_{v})\, dr(\theta, \varphi). \quad (6.24)$$

The second integral in (6.24) is upper bounded by $r(G^{c})$. To bound the first integral we observe that

$$\nu_v(F) = \int \nu^{\theta}_{v}(F)\, dw(\theta) \qquad \text{and} \qquad \hat{\nu}_v(F) = \int \hat{\nu}^{\varphi}_{v}(F)\, d\hat{w}(\varphi).$$

Since $r \in w \nabla \hat{w}$, Property (A.4)(c) implies

$$\int \bar{d}(\nu^{\theta}_{v}, \hat{\nu}^{\varphi}_{v})\, dr(\theta, \varphi) \geqslant \bar{d}(\nu_v, \hat{\nu}_v).$$

Replacing the terms in (6.24) by their bounds gives

$$\int_G E_{\lambda_{\theta,\varphi}}[d_N] \geqslant |\bar{d}(\nu_v, \hat{\nu}_v) - r(G^{c})|^{+},$$

where $|\gamma|^{+} \triangleq \max\{\gamma, 0\}$. Finally, putting this inequality and (6.23) into (6.22) gives

$$E_{\lambda}[d_N] \geqslant |\bar{d}(\nu_v, \hat{\nu}_v) - r(G^{c})|^{+} + \frac{r(G^{c})}{6} \geqslant \frac{\bar{d}(\nu_v, \hat{\nu}_v)}{6},$$

where the last inequality is easy to derive. This is the desired result (6.16) and completes the proof.

(8c) Now we give an example of a pair of $\bar{d}$-continuous channels and ν and $\hat{\nu}$ such that $\bar{D}(\nu, \hat{\nu}) > D(\nu, \hat{\nu})$. Let L be an even integer, let $\underline{u}$ be the sequence $0\,1\,1\,1 \cdots 1$ of length L, and let ν and $\hat{\nu}$ be binary deterministic channels $(A_1 = B_1 = \{0, 1\})$ such that if $X_k^{k+L-1} = \underline{u}$, then $Y_k^{k+L-1} = 0\,1\,0\,1 \cdots 0\,1$ and $\hat{Y}_k^{k+L-1} = 1\,0\,1\,0 \cdots 1\,0$, while if $X_{k-i}^{k-i+L-1} \neq \underline{u}$ for all $i \in \{0, 1,..., L-1\}$, then $Y_k = \hat{Y}_k = X_k$. Notice that ν and $\hat{\nu}$ produce identical outputs where and only where $\underline{u}$ does not occur in the input. Second, since Y_k and $\hat{Y}_k$ are determined by $x_{k-L+1}, x_{k-L+2},..., x_{k+L-1}$, both channels have finite input memory and are therefore $\bar{d}$-continuous (Neuhoff and Shields 1979).

In order to compute $\bar{D}(\nu, \hat{\nu})$ we consider the periodic input sequence v with period N and $v^L = \underline{u}$. For this input the outputs Y_k and $\hat{Y}_k$ disagree for all k. Hence, for any n, $\bar{d}_n(v_v, \hat{v}_v) = 1$. It follows that $\bar{D}(v, \hat{v}) = 1$.

We now show that $D(v, \hat{v}) < 1$. Let α be a stationary source. We shall obtain two bounds for $\bar{d}(\alpha v, \alpha\hat{v})$ by choosing two stationary joinings ω and $\bar{\omega}$ of αv and $\alpha\hat{v}$ and using the inequality

$$\bar{d}(\alpha v, \alpha\hat{v}) \leqslant E_w d_n((XY), (X\hat{Y})), \tag{6.25}$$

which by Property (A.2)(b) holds for any n and any stationary joining ω and αv and $\alpha\hat{v}$.

The first bound is obtained by choosing ω so that $X = \hat{X}$ with probability one. That is, we let

$$\omega(E \times F \times \hat{E} \times \hat{F}) = \int_{E \cap \hat{E}} v_x(F)\, \hat{v}_x(\hat{F})\, d\alpha(x)$$

and observe that ω is stationary and has marginals αv and $\alpha\hat{v}$, respectively. Since ω makes $X = \hat{X}$, it follows that for any n

$$\begin{aligned} \bar{d}(\alpha v, \alpha\hat{v}) &\leqslant E_\omega d_n((XY), (X\hat{Y})) \\ &= E_\omega d_n(Y, \hat{Y}) \\ &= \int E_{v_x\hat{v}_x}[d_n(Y, \hat{Y})]\, d\alpha(x), \end{aligned} \tag{6.26}$$

where $v_x\hat{v}_x$ denotes the product measure on $B \times B$. When $X = \hat{X} = x$, the sequences Y^n and $\hat{Y}^n$ disagree only where $\underline{u}$ occurs, and for each occurrence of $\underline{u}$ there will be L disagreements. Let $N(x^n)$ denote the number of occurrences of $\underline{u}$ in x^n. That is, $N(x^n)$ is the number of integers i, $1 \leqslant i \leqslant n - L + 1$, such that $x_i^{i+L-1} = \underline{u}$. It is easy to see that

$$E_{v_x\hat{v}_x} d_n(Y, \hat{Y}) \leqslant \frac{N(x^n)L}{n} + \frac{L}{n}.$$

Substituting this into (6.26), then using the fact that $E_\alpha[N(x^n)] = (n - L + 1)\, \alpha(\langle \underline{u} \rangle)$ and simplifying gives

$$\bar{d}(\alpha v, \alpha\hat{v}) \leqslant \alpha(\langle \underline{u} \rangle)L + \frac{L}{n}. \tag{6.27}$$

The second bound to $\bar{d}(\alpha v, \alpha\hat{v})$ is obtained by choosing $\bar{\omega}$ so that $\hat{X} = TX$ with probability one. That is, we let

$$\bar{\omega}(E \times F \times \hat{E} \times \hat{F}) = \omega(E \times F \times T^{-1}\hat{E} \times T^{-1}\hat{F})$$

and observe that $\bar{\omega}$ has marginals $\alpha\nu$ and $\alpha\hat{\nu}$, respectively. We then have for any n

$$\bar{d}(\alpha\nu, \alpha\hat{\nu}) \leqslant E_{\bar{\omega}} d_n((XY), (\hat{X}\hat{Y}))$$

$$= \int E_{\nu_x \hat{\nu}_{Tx}} [d_n((XY), (\hat{X}\hat{Y}))] \, d\alpha(x). \tag{6.28}$$

When $\hat{X} = TX = Tx$, then wherever $\underline{u}$ occurs in x, there will be $L - 2$ places where $\hat{X}$ agrees with X. Furthermore in these $L - 2$ places $\hat{Y}$ will agree with Y. Hence the number of places where $(XY)^n$ agrees with $(\hat{X}\hat{Y})^n$ is at least $N(x^n)(L - 2)$, so that

$$E_{\nu_x \hat{\nu}_{Tx}} [d_n((XY), (\hat{X}\hat{Y}))] \leqslant 1 - \frac{N(x^n)(L-2)}{n}.$$

Substituting this into (6.28) and simplifying gives

$$\bar{d}(\alpha\nu, \alpha\hat{\nu}) \leqslant 1 - \alpha(\langle \underline{u} \rangle)(L-2)\frac{n - L + 1}{n}$$

$$\leqslant 1 - \alpha(\langle \underline{u} \rangle)L + \frac{2}{L} + \frac{L}{n} + \frac{2}{Ln}.$$

This inequality together with (6.27) shows that for any α and n

$$\bar{d}(\alpha\nu, \alpha\hat{\nu}) \leqslant \frac{1}{2} + \frac{L}{n} + \frac{2}{L} + \frac{2}{Ln}.$$

It follows that

$$D(\nu, \hat{\nu}) \leqslant \frac{1}{2} + \frac{2}{L},$$

and if $L \geqslant 6$, $D(\nu, \hat{\nu}) < 1$.

(8d) Let ν be a $\bar{d}$-continuous channel. The construction given in (Neuhoff and Shields 1979, Appendix A) provides a distinct channel $\hat{\nu}$ such that $\bar{D}(\nu, \hat{\nu}) = 0$, and it is easy to see that $\hat{\nu}$ is also $\bar{d}$-continuous. Hence $\bar{D}$ is not a metric on the $\bar{d}$-continuous channels and, consequently, neither are the weaker distances.

(9a) The equicontinuity of the mappings $\{\psi_\alpha\}_{\alpha \in \mathbf{B}}$ from $(\mathbf{C}, D_{\mathrm{B}})$ into $(\mathbf{P}, \bar{d})$ follow directly from the definition of D_{B}.

(9b) The example presented in (6c) shows that equicontinuity is lost when D_{B} is replaced by D.

(9c) It follows directly from the definition that the mappings $\{\psi_\alpha\}_{\alpha\in\mathbf{S}}$ from $(\mathbf{C}, D)$ into $(\mathbf{P}, \bar{d})$ are equicontinuous.

(10) Statements (10a) and (10b) are obviously equivalent, as are statements (10c) and (10e). Statement (10b) implies (10c) since $D_{\mathrm{B}} \geqslant D$. Statement (10c) implies (10d) because if $D=0$, then for any stationary source α we have $\alpha\nu = \alpha\hat{\nu}$. Since ν_x and $\hat{\nu}_x$ are each versions of the conditional probability of Y given X, they must be identical for almost all x. Therefore for any $\alpha \in \mathbf{S}$ and any n

$$\int \bar{d}_n(\nu_x, \hat{\nu}_x)\, d\alpha(x) = 0,$$

and this implies $D_{\mathrm{S}}(\nu, \hat{\nu}) = 0$. Finally statement (10d) implies (10b) because $D_{\mathrm{S}} \geqslant D_{\mathrm{B}}$.

Appendix

Properties of the $\bar{d}$ Distance

(A.1) The partition definition of $\bar{d}_n(\alpha, \beta)$ (Ornstein 1973):

Let α, β be measures on A^n, where $A_1 = \{a_1,..., a_J\}$, $J < \infty$. Let (C, λ) be a monatomic measure space. A sequence of partitions $P^{(1)}, P^{(2)},..., P^{(n)}$, where $P^{(i)} = \{P_x^{(i)}: x \in A_1\}$ is said to reflect α if

$$\lambda\left(\bigcap_{i=1}^{n} P_{x_i}^{(i)}\right) = \alpha(\{x^n\}), \qquad \text{all } x^n \in A^n.$$

For any $\{P^{(i)}\}_{i=1}^n$ that reflects α,

$$\bar{d}_n(\alpha, \beta) = \inf_{\{Q^{(i)}\}} \frac{1}{n} \sum_{i=1}^{n} |P^{(i)} - Q^{(i)}|,$$

where the infimum is overall sequence of partitions $\{Q^{(i)}\}_{i=1}^n$ on λ that reflect β and where

$$|P^{(i)} - Q^{(i)}| \triangleq \tfrac{1}{2} \sum_{j\neq k} \lambda(P_j^{(i)} \cap Q_k^{(i)}).$$

(A.2) For N-stationary processes and α and β

(a) $\bar{d}(\alpha, \beta) = \sup_k \bar{d}_{kN}(\alpha, \beta)$.

(b) For any m

$$\bar{d}(\alpha, \beta) = \inf_{\omega} E_\omega[d_{mN}(X, Y)],$$

where the infimum is over all N-stationary measures $\omega \in \alpha \nabla \beta$. The above are straightforward generalizations of the well-known special case where α and β are stationary ($N = 1$) (Ornstein 1973).

(c) If α and β are stationary and ergodic then the infimum over ω can be restricted to stationary ergodic ω (Ornstein 1973).

(A.3) Let α and β be measures on A^n and let $n = l + m$, then

(a) $n\bar{d}_n(\alpha, \beta) \geqslant l\bar{d}_l(\alpha_1^l, \beta_1^l) + m\bar{d}_m(\alpha_{l+1}^n, \beta_{l+1}^n)$,

(b) equality holds if α is the product of α^l and α_{l+1}^n and β is the product of β^l and β_{l+1}^n.

(A.4) Convexity:

In the proof of Proposition (A.1) of (Neuhoff and Shields 1979) it was shown that for any source α, and channels ν and $\hat{\nu}$, and integer n

(a) $\bar{d}_n(\alpha\nu, \alpha\hat{\nu}) \leqslant \int \bar{d}_n(\nu_x^{n/}, \hat{\nu}_x^n)\, d\alpha(x)$.

The techniques used to prove the above can also be used to establish (b) and (c).

(b) If α and β are measures on A^n with arbitrary decompositions

$$\alpha = \int \alpha_\theta \, d\omega(\theta) \qquad \text{and} \qquad \beta = \int \beta_\theta \, d\omega(\theta)$$

then

$$\bar{d}_n(\alpha, \beta) \leqslant \int \bar{d}_n(\alpha_\theta, \beta_\theta)\, dw(\theta).$$

(c) If α and β are N-stationary processes on A with arbitrary decompositions

$$\alpha = \int \alpha_\theta \, dw_\alpha(\theta) \qquad \text{and} \qquad \beta = \int \beta_\varphi \, dw_\beta(\varphi),$$

then for any measure $r \in w_\alpha \nabla w_\beta$,

$$\bar{d}(\alpha, \beta) \leqslant \int \bar{d}(\alpha_\theta, \beta_\varphi)\, dr(\theta, \varphi).$$

(A.5) If ν and $\hat{\nu}$ are stationary, then

$$\bar{d}(\nu_{Tx}, \hat{\nu}_{Tx}) = \bar{d}(\nu_x, \hat{\nu}_x) \qquad \text{all } x.$$

As the proof is straightforward, we omit it.

References

Bailey, D. (1976), "Sequential Schemes for Classifying and Predicting Ergodic Processes," Ph.D. thesis, Stanford University.

Gallager, R. G. (1968), "Information Theory and Reliable Communication," Wiley, New York.

Gray, R. M., and Davisson, L. D. (1974), The ergodic decomposition of discrete starionary sources, *IEEE Trans. Inform. Theory* **IT-20**, 625–636.

Gray, R. M., and Ornstein, D. S. (1979), Block coding for discrete stationary $\bar{d}$-continuous channels, *IEEE Trans. Inform. Theory* **IT-25**, 292–306.

Neuhoff, D. L., and Shields, P. C. (1979), Channels with almost finite memory, *IEEE Trans. Inform. Theory* **IT-25**, 440–447.

Neuhoff, D. L., and Shields, P. C. (1982a), Indecomposable finite state channels and primitive approximation, *IEEE Trans. Inform. Theory* **IT-28**, 11–18.

Neuhoff, D. L. and Shields, P. C. (1982b), Channel entropy and primitive approximation, *Ann. Probab.* **10**, 188–198.

Ornstein, D. S. (1973), An application of ergodic theory to probability theory, *Ann. Probab.* **1**, 43–65.

Rohlin, V. A. (1962), On the fundamental ideas of measure theory, *Mat. Sb.* **25**, 107–150; Amer. Math. Soc. Transl., Ser. 1, pp. 1–54.