The Annals of Probability
1975, Vol. 3, No. 2, 315-328

A GENERALIZATION OF ORNSTEIN'S $\bar{d}$ DISTANCE WITH APPLICATIONS TO INFORMATION THEORY[1]

BY ROBERT M. GRAY, DAVID L. NEUHOFF AND PAUL C. SHIELDS

Stanford University, Universiy of Michigan and University of Toledo

Ornstein's $\bar{d}$ distance between finite alphabet discrete-time random processes is generalized in a natural way to discrete-time random processes having separable metric spaces for alphabets. As an application, several new results are obtained on the information theoretic problem of source coding with a fidelity criterion (information transmission at rates below capacity) when the source statistics are inaccurately or incompletely known. Two examples of evaluation and bounding of the process distance are presented: (i) the $\bar{d}$ distance between two binary Bernoulli shifts, and (ii) the process distance between two stationary Gaussian time series with an alphabet metric $|x - y|$.

1. Introduction. In an invited paper in *The Annals of Probability*, Ornstein (1973) discussed the concept of the $\bar{d}$ distance between finite alphabet, discrete-time random processes that he had developed earlier in several papers referenced therein. In a discussion of Ornstein's paper, Krengel (1973) predicted that much of Ornstein's work would lead to new methods and results in Shannon information theory. In this paper we present some initial progress in this direction by demonstrating that a natural generalization of the $\bar{d}$ distance can be used to obtain results on source coding with a fidelity criterion when the source statistics are inaccurately or incompletely known. This provides lower bounds to attainable average fidelity when trying to communicate a source with inaccurately or incompletely known statistics across a channel with capacity possibly less than the source entropy (which may be infinite). As examples, the process distance is evaluated or bounded for two special cases:

(i) the $\bar{d}$ distance between two binary Bernoulli shifts, and
(ii) the generalized distance between two stationary Gaussian time series with alphabet metric $|x - y|$.

2. The $\bar{\rho}$ distance. Let the alphabet A be a separable complete metric space with metric ρ. Let $\mathscr{A}$ be the σ-field generated by the open sets of A. Since A is separable metric space, $\mathscr{A}$ thus defined is a separable σ-field with some

Received November 5, 1973; revised March 18, 1974.

[1] This was supported in part by NSF Grant GK-31630, NSF Grant GJ776, and by the Joint Service Program at Stanford Electronic Laboratories under U.S. Navy Contract N00014-67-A-0112-0044. Neuhoff was at Stanford University during the course of this research.

AMS 1970 *subject classifications.* Primary 60G35, 94A15; Secondary, 94A05.

Key words and phrases. $\bar{d}$ and $\bar{\rho}$ distance, stationary time series, source coding with a fidelity criterion, Gaussian time series.

countable set of generators $\mathscr{G} = \{\alpha_i\}$. For each n define the measurable space

$$(A^n, \mathscr{A}^n) = \mathsf{X}_{k=0}^{n-1} (A_k, \mathscr{A}_k) ,$$

where A^n is the Cartesian product of exemplars of A_k of A and $\mathscr{A}^n$ is the smallest σ-field containing all sets of the form $\mathsf{X}_{i=0}^{n-1} B_i$, $B_i \in \mathscr{G}^{n-1}$. The metric on $A^n \times A^n$ is defined as

$$\rho_n(x^n, y^n) = n^{-1} \textstyle\sum_{i=0}^{n-1} \rho(x_i, y_i)$$

where $x^n = (x_0, \cdots, x_{n-1}) \in A^n$. Define the sequence measurable space $(\Sigma, \mathscr{S}) = (A^\infty, \mathscr{A}^\infty)$, i.e., Σ is the space of doubly infinite sequences of elements of A and $\mathscr{S}$ is the smallest σ-field containing all of the cylinders of the form $C = \cdots \times A_{i-2} \times A_{i-1} \times C_i \times C_{i+1} \times \cdots \times C_{i+n-1} \times A_{i+N} \times \cdots$, where each $C_i \in \mathscr{G}_i$. Denote the countable set of all such cylinders by $\mathscr{C}$. As does Blumenthal (1973) in his discussion of Ornstein's paper, we consider sequence space rather than an underlying probability space as this is more natural for information theoretic applications.

Let μ_θ be shift invariant (stationary) measures on $(\Sigma, \mathscr{S})$ for all values of an index θ so that $\{(\Sigma, \mathscr{S}, \mu_\theta)\}$ is a family of probability spaces. We shall make use of the fact that μ_θ is uniquely specified by its values on $\mathscr{G}$. Let $\mu_\theta{}^n$ denote the restriction μ_θ to $(A^n, \mathscr{A}^n)$. For elements $x = (\cdots, x_{-1}, x_0, x_1, \cdots) \in \Sigma$ let $X_n(x) = x_n$ denote the projection of x onto the coordinate space A_n and let T denote the shift on Σ, i.e., $(Tx)_n = x_{n+1}$. The sequence of random variables $X_n(x) = X_0(T^n x)$ described by $(\Sigma, \mathscr{S}, \mu_\theta)$ is then a stationary discrete time random process and will be abbreviated by $[A, \mu_\theta]$.

Let $[A, \mu_\theta]$ and $[A, \mu_\varphi]$ be two processes with corresponding sequences of random variables $\{X_n\}$ and $\{Y_n\}$, respectively. The processes are considered identical if $\mu_\theta(B) = \mu_\varphi(B)$, all $B \in \mathscr{C}$. Define the $\bar{\rho}$ distance between the two processes, $\bar{\rho}(\theta, \varphi)$, as follows:

$$\bar{\rho}(\theta, \varphi) = \sup_n \bar{\rho}_n(\theta, \varphi) ,$$

$$\bar{\rho}_n(\theta, \varphi) = \inf_{p^n \in \mathscr{P}_n} E_{p^n}[\rho_n(X^n, Y^n)] ,$$

where

$$E_{p^n}[\rho_n(X^n, Y^n)] = \textstyle\int_{A^n \times A^n} \rho_n(x^n, y^n) dp^n(x^n, y^n) ,$$

and $\mathscr{P}_n$ is the class of joint measures p^n on $(A^n, \mathscr{A}^n)^2$ yielding $\mu_\theta{}^n$ and $\mu_\varphi{}^n$ as marginals, i.e.,

$$\mathscr{P}_n = \{p^n : p^n(B \times A^n) = \textstyle\int_{B \times A^n} dp^n(x^n, y^n) = \mu_\theta{}^n(B),$$
$$p^n(A^n \times B) = \textstyle\int_{A^n \times B} dp^n(x^n, y^n) = \mu_\varphi{}^n(B), \text{ all } B \in \mathscr{A}^n\} .$$

Kailath (1973) has pointed out that $\bar{\rho}_n(\theta, \varphi)$ is the Vasershtein distance (Vasershtein (1969)) between the measures $\mu_\theta{}^n$ and $\mu_\varphi{}^n$ with metric ρ_n on A^n.

The $\bar{\rho}$ distance between two processes is thus a measure of how closely we can "fit" the two processes together in a random manner according to a given metric. We shall shortly see that the various interpretations of the $\bar{d}$ distance have corresponding interpretations of the $\bar{\rho}$ distance. We next state and prove several straightforward properties of the $\bar{\rho}$ distance.

PROPERTIES OF $\bar{\rho}$:

(i) $\lim_{n\to\infty} \bar{\rho}_n(\theta, \varphi)$ exists and equals $\sup \bar{\rho}_n(\theta, \varphi)$;

(ii) $\bar{\rho}$ is a metric;

(iii) if A_0 is finite and $\rho(k, j) = 1 - \delta_{k,j}$, where $\delta_{k,j}$ is the Kronecker δ, then $\bar{\rho}(\theta, \varphi) = \bar{d}(\theta, \varphi)$;

(iv) if μ_θ and μ_φ both yield i.i.d. random processes, then $\bar{\rho}(\theta, \varphi) = \bar{\rho}_1(\theta, \varphi)$;

(v) $\bar{\rho}_1(\theta, \varphi) \leqq \bar{\rho}(\theta, \varphi) \leqq \int \rho \, d\mu_\theta^1 \, d\mu_\varphi^1$.

PROOF.

(i) Let p^N yield $\bar{\rho}_N$ (or arbitrary close if the minimum is not achieved). For any $n < N$ define

$$p^n(B) = p^N(B \times A^{2(N-n)}), \qquad B \in \mathscr{A}^{2n},$$

$$p^{N-n}(C) = p^N(C \times A^{2n}), \qquad C \in \mathscr{A}^n.$$

We have immediately that $p^n \in \mathscr{P}_n$ and $p^{2(N-n)} \in \mathscr{P}_{N-n}$ and therefore

$$\begin{aligned} N\bar{\rho}_N(\theta, \varphi) &= nE_p n\rho_n(X^n, Y^n) + (N-n)E_p N - n\rho_{N-n}(X^{N-n}, Y^{N-n}) \\ &\leqq n\bar{\rho}_n(\theta, \varphi) + (N-n)\bar{\rho}_{N-n}(\theta, \varphi) \end{aligned}$$

which implies that the limit of $\bar{\rho}_n$ exists and equals the supremum (cf. Gallager (1968) page 112).

(ii) Parallel to the $\bar{d}$ distance, the fact that ρ is a metric and that the above limit exists implies that $\bar{\rho}$ satisfies the triangle inequality. $\bar{\rho}(\theta, \varphi)$ is obviously symmetric in its arguments and nonnegative. If $\bar{\rho}(\theta, \varphi) = 0$, then $\bar{\rho}_n(\theta, \varphi) = 0$ for all n which implies that there exist $p^n(x^n, y^n : x^n \neq y^n) = 0$ such that for all $B \in \mathscr{A}^n$ we have $\mu_\theta{}^n(B) = p^n(B \times A^n)$ and $\mu_\varphi{}^n(B) = p^n(A^n \times B)$.
This is only possible, however, if $\mu_\theta{}^n(B) = \mu_\varphi{}^n(B)$, all $B \in \mathscr{A}^n$. Since this holds for all n, $[A, \mu_\theta]$ and $[A, \mu_\varphi]$ are the same random process. Hence $\bar{\rho}$ is a metric.

(iii) This follows from the fact that for each n, $\bar{\rho}_n(\theta, \varphi) = \bar{d}(\{P_i\}_1^n, \{Q_i\}_1^n)$ as defined by Ornstein (1973).

(iv) Let p^1 yield $\bar{\rho}_1$ (or k arbitrarily close) and let p^n be the product measure of np^1's. This yields an upper bound to $\bar{\rho}_n$ since $p^n \in \mathscr{P}_n$ if $\mu_\theta{}^n$ and $\mu_\varphi{}^n$ are themselves product measures of $\mu_\theta{}^1$ and $\mu_\varphi{}^1$. Thus

$$\bar{\rho}_n(\theta, \varphi) \leqq \int \rho_n \, dp^n = n^{-1} \sum_{i=0}^{n-1} \int \rho \, dp^1 = \bar{\rho}_1(\theta, \varphi).$$

Since by definition $\bar{\rho} = \sup \bar{\rho}_n$, $\bar{\rho} = \bar{\rho}_1$.

(v) The left-hand inequality follows from the definition. The right-hand inequality follows from the test measure $p^n = \mu_\theta{}^n \times \mu_\varphi{}^n$.

We next present two alternative definitions of distances between processes.

Recall that $\mathscr{C}$ is a countable basis for $\mathscr{S}$. Call a string $x \in \Sigma$ θ-representative if

$$\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} I_C(T^{-k}x) = \mu_\theta(C), \qquad \text{all } C \in \mathscr{C},$$

where I_C is the indicator function. Denoting the collection of all θ-representative strings by Σ_θ, we have from the ergodic theorem that for ergodic processes $\mu_\theta(\Sigma_\theta) = 1$. Define the $\bar{\rho}'$ distance between two ergodic processes $[A, \mu_\theta]$ and

$[A, \mu_\varphi]$ as

$$\bar{\rho}'(\theta, \varphi) = \inf_{x \in \Sigma_\theta, y \in \Sigma_\varphi} \limsup_{n \to \infty} n^{-1} \sum_{i=0}^{n-1} \rho(x_i, y_i) .$$

The $\bar{\rho}'$ distance tells us the smallest necessary amount by which we must change a representative string of one process to make it look like a representative string of the other process.

Let $\mathscr{E}_{\theta,\varphi}$ be the set of all jointly stationary measures on $(\Sigma, \mathscr{S})^2$ having μ_θ and μ_φ as marginals. Define

$$\bar{\rho}''(\theta, \varphi) = \inf_{p \in \mathscr{E}_{\theta,\varphi}} E_p[\rho(X_0, Y_0)] .$$

The $\bar{\rho}''$ distance tells how closely two jointly stationary processes can be fit together at a single time.

THEOREM 1. *$\bar{\rho}''(\theta, \varphi) = \bar{\rho}(\theta, \varphi)$, and, if $[A, \mu_\theta]$ and $[A, \mu_\varphi]$ are each ergodic, then*

$$\bar{\rho}(\theta, \varphi) = \bar{\rho}'(\theta, \varphi) = \bar{\rho}''(\theta, \varphi) .$$

Thus the $\bar{\rho}$ distance tells us

(i) how much we must change an entire representative sample function of one process for it to be confused with a representative of the other process, and

(ii) how closely we can force the two processes to resemble each other at a single time instant if the two processes are generated by a single stationary two dimensional process. The proof of Theorem 1 is an adaptation of the corresponding results for the $\bar{d}$ distance (Ornstein (1973)) and is relegated to an appendix.

As suggested by Kailath (1973), $\bar{\rho}''$ can be interpreted as the Vasershtein distance (Vasershtein (1969)) between μ_θ and μ_φ with metric $\limsup_{n \to \infty} n^{-1} \sum_{i=0}^{n-1} \rho(x_i, y_i)$ on $\sum$ and with the constraint that the class of joint probability distributions on $(\Sigma, \mathscr{S})$ be stationary. Hence Theorem 1 demonstrates the equivalence of the $\bar{\rho}$ distance with the appropriate Vasershtein distance.

We note that the $\bar{\rho}$ distance as defined here extends naturally to continuous time random processes with sums replaced by integrals.

3. Applications to information theory. We begin with a summary of the relevant definitions and theorems of distortion-rate theory (theory of source coding subject to a fidelity criterion) in the appropriate notation. The theory had its origins in Shannon's classic 1948 paper, and was further developed by Shannon (1959). Modern measure theoretic expositions may be found in Berger (1971) and Gray and Davisson (1974).

For an integer N a codebook C_N is a collection of $\|C_N\|$ N-tuples $\{y_i^N\}_{i=1,\cdots,\|C_N\|}$ drawn from $\hat{A}^N = \times_{k=0}^{N-1} \hat{A}_k$, where $\hat{A}_k$ are replicas of the so-called "available reproducing alphabet" $\hat{A}$ (usually $\hat{A} \subseteq A$). We assume that ρ is a metric on $A \cup \hat{A}$. Each successive source N-tuple x^N is encoded into the codeword $y^N \in C_N$ minimizing $\rho_N(x^N, y^N)$. The resulting codeword is denoted by $\hat{x}(x^N)$. Define

$$\rho_N(x^N \mid C_N) = \min_{y^N \in C_N} \rho_N(x^N, y^N) = \rho_N(x^N, \hat{x}(x^N)) .$$

The average distortion of a codebook C_N used to encode a source $[A, \mu_\theta]$ is

given by

$$\rho(C_N \mid \theta) =_\Delta E_{\mu\theta}[\rho_N(X^N \mid C_N)] =_\Delta \int_{A^N} \rho_N(x^N \mid C_N)\, d\mu_\theta{}^N(x^N) .$$

The rate $R(C_N)$ of a codebook is defined by $R(C_N) = N^{-1} \ln ||C_N||$. Let $\mathscr{E}(N, R)$ denote the collection of all block length N codebooks having rate R or less. The quantity of interest is the smallest attainable average distortion using codebooks of a given block length and rate to encode successive source blocks, i.e.,

$$\delta_\theta(R, N) = \inf_{C_N \in \mathscr{E}(N,R)} \rho(C_N \mid \theta) .$$

Also of interest is the minimum attainable average distortion over all rate R codes:

$$\delta_\theta(R) = \inf_N \delta_\theta(R, N) .$$

It is easily shown that the limit of $\delta_\theta(R, N)$ as $N \to \infty$ exists and equals the infimum. Coupling this definition with Shannon's channel coding theorem we can interpret $\delta_\theta(R)$ as follows: If we attempt to send information about a source $[A, \mu_\theta]$ across a channel of capacity C, then the minimum attainable average distortion between the source and the channel output using block coding is $\delta_\theta(C)$. (This is one version of the Information Transmission Theorem, Gallager (1968) page 449.)

Information theory provides a means of evaluating $\delta_\theta(R)$ via standard convex minimization techniques. Let q^n be a conditional probability measure on $\hat{\mathscr{A}}^n$, a σ-algebra of subsets of $\hat{A}^n$, given events in $\mathscr{A}^n$. Let p^n denote the joint probability measure induced on $\mathscr{A}^n \times \hat{\mathscr{A}}^n$ by μ^n and q^n, i.e., for any set $G \in \mathscr{A}^n \times \hat{\mathscr{A}}^n$

$$p^n(G) = \int q^n(x, G_x)\, d\mu^n(x)$$

where $G_x = \{y : (x, y) \in G\}$. Following Pinsker (1960) or Berger (1971), define the average mutual information I_n of the joint probability space $(A^n \times \hat{A}^n, \mathscr{A}^n \times \hat{\mathscr{A}}^n, p)$ by

$$I_n = I(\mu^n, q^n) = \sup \sum_{i=1}^{\infty} p^n(G_i \times B_i) \log \frac{p^n(G_i \times B_i)}{\mu^n(G_i)\nu^n(B_i)}$$

where $\nu^n(B_i) = p^n(A^n \times B_i)$ and the supremum is taken over all partitions of $A^n \times \hat{A}^n$ into countably many rectangles $G_i \times B_i$; $G_i \in \mathscr{A}^n$, $B_i \in \hat{\mathscr{A}}^n$. We define the distortion-rate function of the source $[A, \mu_\theta]$ and distortion measure ρ as follows: For $R \in [0, \infty]$, define $Q_n(R)$ as the class of probability measures q^n such that $n^{-1} I(\mu_\theta{}^n, q^n) \leqq R$. Let

$$D_{\theta,n}(R) = \inf_{q^n \in Q_n(R)} \int \rho_n(x^n, y^n)\, dp^n(x^n, y^n) ,$$

with $D_{\theta,n}(R) = +\infty$ if $Q_n(R)$ is empty, and define

$$D_\theta(R) = \lim_{n \to \infty} D_{\theta,n}(R) .$$

It is easily shown that for stationary sources $D_\theta(R)$ exists and is a nonnegative, monotonic, decreasing convex $\cup$ function over the interval in which it is finite. The basic source coding theorem (Shannon (1949), Gallager (1968), Berger (1971)) can be stated as follows (Gray, Davisson (1974)):

THEOREM 2. SOURCE CODING THEOREM. *Given an ergodic source* $[A, \mu_\theta]$ *with distortion-rate function* $D_\theta(R)$, *assume there exists a letter* $\hat{x}^*$ *for which*

$$E_{\mu_\theta}[\rho_1(x, \hat{x}^*)] =_\Delta \int \rho_1(x, \hat{x}^*)\, d\mu_\theta(x) \leqq \rho^* < \infty ,$$

then $D_\theta(R) = \delta_\theta(R)$ *and hence there exists a sequence of codes* C_N *such that*

$$\lim_{N\to\infty} \rho(C_N \mid \theta) = D_\theta(R) = \delta_\theta(R) .$$

The generalization to stationary nonergodic sources via the ergodic decomposition can be found in Gray and Davisson (1974).

One of the major shortcomings of the theory is that in practice source statistics are not known precisely. The calculation of $D_\theta(R)$, however, requires precise knowledge of the measure μ_θ. Two questions thus naturally arise:

(i) If a sequence of codes is designed for a source model $[A, \mu_\theta]$, but the actual source is $[A, \mu_\varphi]$, how much performance might be lost due to this mismatch?

(ii) What is the minimum attainable average distortion using fixed rate block coding if we must encode an unknown member of a class of sources?

The first problem has not been treated in the information theoretic literature to our knowledge. We shall see that the $\bar{\rho}$ distance provides a simple solution. The second problem is that of universal coding subject to a fidelity criterion and has been studied by Dobrushin (1963), Sakrison (1969), Ziv (1972), Gray and Davisson (1974), and Neuhoff, Gray, and Davisson (1975). Here, the $\bar{\rho}$ distance provides a new intuitive method that provides both some new results and simplified proofs of results similar to known results.

The solution to the first problem and the basic method for attacking the second is given by the following easy result.

THEOREM 3. *Given two stationary sources* $[A, \mu_\theta]$ *and* $[A, \mu_\varphi]$, *any integer* N, *and any codebook* C_N, *then*

$$|\rho(C_N \mid \theta) - \rho(C_N \mid \varphi)| \leqq \bar{\rho}(\theta, \varphi) .$$

PROOF. Let p^N yield $\bar{\rho}_N$ (or be arbitrarily close), then

$$\begin{aligned}\rho(C_N \mid \varphi) &= \int_{A^N} d\mu_\varphi{}^N(y^N)\rho(y^N \mid C_N) = \int_{A^N \times A^N} d\rho^N(x^N, y^N)\rho(y^N \mid C_N) \\ &\leqq \int_{A^N \times A^N} d\rho^N(x^N, y^N)[\rho_N(x^N, y^N) + \rho(x^N \mid C_N)] \\ &= \bar{\rho}_N(\theta, \varphi) + \rho(C_N \mid \theta) \leqq \rho(C_N \mid \theta) + \bar{\rho}(\theta, \varphi) .\end{aligned}$$

Interchanging the roles of θ and φ yields the theorem.

The above simple proof demonstrates the natural application of the $\bar{\rho}$ metric to information theoretic problems involving approximation.

If a sequence of codes C_N is designed for $[A, \mu_\theta]$ so that $\rho(C_N \mid \theta) \to_{N\to\infty} \delta_\theta(R)$, then the above theorem states that if we apply this sequence to a source $[A, \mu_\varphi]$ we will have limiting average distortion

$$\lim_{N\to\infty} \rho(C_N \mid \varphi) \leqq \lim_{N\to\infty} \rho(C_N \mid \theta) + \bar{\rho}(\theta, \varphi) = \delta_\theta(R) + \bar{\rho}(\theta, \varphi)$$

so that codes will work well for the "true" mismatched source if the $\bar{\rho}$ distance

between actual source and the design model is small. We note that this conclusion does not follow from closeness in the vague topology (distribution metric). Since the left-hand side is bound below by $\delta_\theta(R)$, we have immediately the following corollary;

COROLLARY 1. *Given two stationary processes having separable metric spaces for alphabets, then*

$$|\delta_\theta(R) - \delta_\varphi(R)| \leqq \bar{\rho}(\theta, \varphi) .$$

In particular, if the sources are ergodic

$$|D_\theta(R) - D_\varphi(R)| \leqq \bar{\rho}(\theta, \varphi) .$$

The distortion-rate inequality can easily be extended to nonergodic stationary sources using the first inequality and the coding theorem for nonergodic sources (Gray and Davisson (1974)).

We next consider the problem of universal source coding. Consider a class of ergodic sources $\{[A, \mu_\theta]; \theta \in \Lambda\}$. The object is to construct a sequence of codes without knowledge of θ that will work well regardless of the actual source chosen by nature. A sequence of codes C_N is said to be weakly-minimax universal if

$$\lim_{N\to\infty} R(C_N) = R ,$$
$$\lim_{N\to\infty} \rho(C_N \mid \theta) = \delta_\theta(R) , \qquad \text{all } \theta \in \Lambda .$$

The sequence C_N is said to be strongly-minimax if the above limits are uniform in θ, i.e., if given $\varepsilon > 0$ there exists a single N_0 not dependent on θ such that

$$|R(C_N) - R| < \varepsilon ,$$
$$|\rho(C_N \mid \theta) - \delta_\theta(R)| < \varepsilon , \qquad \text{all } \theta \in \Lambda,\ N \geqq N_0 .$$

When actually constructing finite length codes, strongly-minimax universality is more desirable. Weakly-minimax universality is usually easier to demonstrate, however, and provides a good class of codes to search for the stronger variety. Ziv (1972) has shown via a complicated covering argument that under tighter restrictions on the source alphabet that weakly-minimax universal codes always exist for arbitrary classes of stationary sources. We prove a parallel result for the more general alphabet of separable metric spaces but less general measures. Thus our result reinforces Ziv's and neither subsumes the other. The point here, however, is that the $\bar{\rho}$ metric provides a much simpler and more direct and intuitive proof. As a corollary we then give sufficient conditions for the existence of strongly-minimax universal codes. This is the only such result known to the authors.

THEOREM 4. UNIVERSAL CODING. *If the class $\{[A, \mu_\theta]; \theta \in \Lambda\}$ is a separable metric space under $\bar{\rho}$ and if there exists an $\hat{x}^*$ such that*

$$E_\theta \rho(x^1, \hat{x}^*) \leqq \rho_\theta{}^* < \infty , \qquad \textit{all } \theta \in \Lambda ,$$

then there exist weakly minimax universal codes.

PROOF. Let $\{\theta_k\}_{k=1}^{\infty}$ index a countably dense (in the $\bar{\rho}$ distance) subset of processes. Given $\varepsilon > 0$, let $\{V_k\}$ be spheres of radius $\varepsilon/3$ about θ_k. Construct a subsequence of codes as follows: For each $K = 1, 2, 3, \cdots$ choose $N(K)$ large enough so that there exist rate R codes $C_{N(K)}(k)$, $k = 1, \cdots, K$, such that

$$\rho(C_{N(K)}(k) \mid \theta_k) \leqq \delta_{\theta_k}(R) + \varepsilon/3$$

and $N(K) \geqq K$.

Form the union codebook

$$C_{N(K)} =_{\Delta} \bigcup_{k=1}^{K} C_{N(K)}(k) .$$

The union codebook is then coupled with the usual encoding rule (find the "best" codeword in $C_{N(K)}$ for any given source vector) providing a block length $N(K)$ code with rate

$$R(C_{N(K)}) = R + N(K)^{-1} \ln K \leqq R + K^{-1} \ln K .$$

Since the "best" codeword in $C_{N(K)}$ for any particular source block can be no worse than the best codeword in any particular subcode $C_{N(K)}(k)$, we have for any θ that

$$\rho(C_{N(K)} \mid \theta) \leqq \sum_{k=1}^{K} \rho(C_{N(K)}(k) \mid \theta) I_{V_k}(\theta) + \rho_{\theta}^{*}[I_{\Lambda}(\theta) - I_{\bigcup_{k=1}^{K} V_k}(\theta)]$$

where $I_A(\theta)$ is the indicator function of the set A. Since the V_k are spheres about a dense subset of Λ, $\Lambda = \bigcup_{k=1}^{\infty} V_k$ and we have from Theorem 3 and its corollary that

$$\begin{aligned} \lim_{K\to\infty} \rho(C_{N(K)} \mid \theta) &\leqq \lim_{K\to\infty} \sum_{k=1}^{K} [\rho(C_{N(K)}(k) \mid \theta_k) + \varepsilon/3] I_{V_k}(\theta) \\ &\leqq \lim_{K\to\infty} \sum_{k=1}^{K} [\delta_{\theta_k}(R) + 2\varepsilon/3] I_{V_k}(\theta) \\ &\leqq \lim_{K\to\infty} \sum_{k=1}^{K} [\delta_{\theta}(R) + \varepsilon] I_{V_k}(\theta) = \delta_{\theta}(R) + \varepsilon . \end{aligned}$$

Since ε is arbitrary and since a subsequence of codes $C_{N(K)}$ yields a sequence C_N with the same properties (Gray and Davisson (1974)), the theorem is proved.

Shields (1974) has shown that the class of all B-processes with a fixed alphabet is separable under the $\bar{d}$ distance and that the class of k state Markov chains is separable under the $\bar{d}$ distance. The class of all independent, identically distributed (i.i.d.) processes with a fixed separable metric space is easily seen to be separable (put rational probabilities on countable generators). Separability of the space of Gaussian time-series with a magnitude alphabet metric follows from the formula of the next section and the separability of $L_1(-\pi, \pi)$ under $\|\cdot\|_1$.

From the proof of the weakly-minimax universal coding theorem we have the following obvious corollary.

COROLLARY 2. *If the space of processes* $\{[A, \mu_\theta]; \theta \in \Lambda\}$ *is compact or totally bounded under* $\bar{\rho}$, *then there exist strongly minimax universal codes.*

The class of all i.i.d. processes with a given finite alphabet is compact under the $\bar{d}$ distance. Shields (1974) has shown that given $\delta > 0$, the class of finite alphabet Markov chains for which any state is reachable in a specified finite

number of steps with probability $\geqq \delta$ is compact. From the results of Section 4 it follows that any class of stationary Gaussian processes with power spectral densities in a compact subset of L_1 is compact. Finally, any finite collection of processes is clearly compact. It would be of interest to obtain more general classes of $\bar{\rho}$-compact classes of random processes.

The $\bar{\rho}$ distance has also proved useful in providing a new definition of Shannon's distortion-rate function and interpreting the source coding theorem (Gray, Neuhoff and Omura (1975)) and in proving a source coding theorem using nonblock codes (Gray, Neuhoff, and Ornstein (1975)).

4. Evaluation of $\bar{\rho}$.

EXAMPLE 1. We first consider the simplest possible example: the $\bar{d}$ distance between two i.i.d. binary sequences with bias p_1 and p_2. This example is also developed in a different way by England and Shields (1974) and is presented here for completeness. Since the processes are i.i.d., $\bar{\rho} = \bar{\rho}_1$ and hence,

$$\bar{\rho}(\theta, \varphi) = \inf_{q(k,j)} : \sum_{k \neq j} q(k, j)$$
$$\sum_{j=0}^{1} q(1, j) = p_2 , \qquad \sum_{k=0}^{1} q(k, 1) = p_2 .$$

A LaGrange Minimization immediately yields $\bar{\rho}(\theta, \varphi) = |p_1 - p_2|$. An interpretation is as follows: if $p_2 \geqq p_1$ and x is p_2-representative, then change a percentage $p_2 - p_1$ of the ones in x to zeroes and the resulting string will be p_1-representative.

EXAMPLE 2. Consider two real stationary Gaussian random time series: Let A be the real line and let $[A, \mu_\theta]$ be a sequence $\{X_n\}$ of zero mean Gaussian random variables described by the correlation function $R_\theta(k) = EX_j X_{j+k}$ and power spectral density (discrete Fourier transform) $f_\theta(\lambda) = \sum_{k=-\infty}^{\infty} R_\theta(k) e^{ik\lambda}$. Similarly, let $[A, \mu_\varphi]$ be a zero mean Gaussian time series $\{Y_n\}$ with correlation $R_\varphi(k)$ and $f_\varphi(\lambda)$. Consider two measures of distance on A: $\rho(x^1, y^1) = |x^1 - y^1|$ and $\rho^*(x^1, y^1) = (x^1 - y^1)^2$. The first measure ρ is a metric and hence all of the previous results and applications apply. The second measure ρ^* is not a metric, but is considered since it is one of the most popular measures of "distortion" in information theory and since all of the properties of the resulting $\bar{\rho}^*$ "distance" remain valid except for the triangle inequality (and therefore the applications described in Section 3). Since other uses of $\bar{\rho}$ do not require the metric property (e.g., Gray, Neuhoff and Omura (1975), and Gray, Neuhoff, and Ornstein (1975)), this non-metric "distance" is presented for comparison and completeness. For simplicity consider the $\bar{\rho}''$ definition and define $\bar{\rho}(\theta, \varphi) = \inf E_p\{|X^1 - Y^1|\}$ and $\rho^*(\theta, \varphi) = \inf E_p\{(X^1 - Y^1)^2\}$, where both infimums are over $\xi_{\theta,\varphi}$. For any stationary $p \in \xi_{\theta,\varphi}$ yielding a pair process $\{X_n, Y_n\}$, define the cross-correlation function $R_{\theta\varphi}(k) = E_p(X_n Y_{n+k})$ and its Fourier transform $f_{\theta\varphi}(\lambda) = \sum R_{\theta\varphi}(k) e^{ik\lambda}$, the so-called cross-spectral density. From Rozanov (1967) page 19, we have for any p that $|f_{\theta\varphi}(\lambda)|^2 \leqq f_\theta(\lambda) f_\varphi(\lambda)$.

Since X^1 and Y^1 are zero-mean Gaussian random variables, $Z = X^1 - Y^1$ is

also a zero-mean random variable with variance σ^2 given by

$$\begin{aligned}\sigma^2 &= E_p Z^2 = R_\theta(0) + R_\varphi(0) - 2R_{\theta\varphi}(0)\\ &= (2\pi)^{-1}\int_{-\pi}^{\pi}[f_\theta(\lambda) + f_\varphi(\lambda) - 2f_{\theta\varphi}(\lambda)]\,d\lambda\\ &\geqq (2\pi)^{-1}\int_{-\pi}^{\pi}[f_\theta(\lambda) + f_\varphi(\lambda) - 2(f_\theta(\lambda)f_\varphi(\lambda))^{\frac{1}{2}}]\,d\lambda =_\Delta \sigma^2 .\end{aligned}$$

We therefore have for any stationary $p \in \xi_{\theta\varphi}$ that $E_p\{(X^1 - Y^1)^2\} = \sigma_z^2 \geqq \sigma^2$ and hence $\bar{\rho}^*(\theta, \varphi) \geqq \sigma^2$.

Since the matrix

$$\begin{bmatrix} f_\theta(\lambda) & (f_\theta(\lambda)f_\varphi(\lambda))^{\frac{1}{2}} \\ (f_\theta(\lambda)f_\varphi(\lambda))^{\frac{1}{2}} & f_\varphi(\lambda)\end{bmatrix}$$

is nonnegative definite, there exists a stationary Gaussian pair process $\{X_n, Y_n\}$ with $f_{\theta\varphi}(\lambda) = (f_\theta(\lambda)f_\varphi(\lambda))^{\frac{1}{2}}$ from Section 1.9 of Rozanov (1967) and therefore we have for the resulting p that $\sigma_z^2 = \sigma^2$ and

$$\begin{aligned}E_p[|X^1 - Y^1|] &= (2/\pi)^{\frac{1}{2}}\sigma \geqq \bar{\rho}(\theta, \varphi)\\ E_p\{(X^1 - Y^1)^2\} &= \sigma^2 \geqq \bar{\rho}^*(\theta, \varphi)\end{aligned}$$

yielding the following Theorem.

THEOREM 5. *If* $[A, \mu_\theta]$ *and* $[A, \mu_\varphi]$ *are stationary Gaussian time series with spectral densities* f_θ *and* f_φ, *respectively, and if* $\rho(x^1, y^1) = |x^1 - y^1|$ *and* $\rho^*(x^1, y^1) = (x^1 - y^1)^2$, *then*

$$\begin{aligned}(2\pi)^{\frac{1}{2}}|\sigma_\theta - \sigma_\varphi| \leqq \bar{\rho}(\theta, \varphi) &\leqq \pi^{-1}\{\int_{-\pi}^{\pi}|f_\theta(\lambda)^{\frac{1}{2}} - f_\varphi(\lambda)^{\frac{1}{2}}|^2\,d\lambda\}^{\frac{1}{2}}\\ \bar{\rho}^*(\theta, \varphi) &= (2\pi)^{-1}\int_{-\pi}^{\pi}|f_\theta(\lambda)^{\frac{1}{2}} - f_\varphi(\lambda)^{\frac{1}{2}}|^2\,d\lambda .\end{aligned}$$

The lower bound to $\bar{\rho}$ follows from the fact that $\bar{\rho} \geqq \bar{\rho}_1$, which has been evaluated by Vallender (1973). If both processes are i.i.d., then $\bar{\rho}(\theta, \varphi) = (2\pi)^{\frac{1}{2}}|\sigma_\theta - \sigma_\varphi|$.

Note that $\bar{\rho}^*(\theta, \varphi)$ is the square of the $L_2(-\pi, \pi)$ distance between the square roots of the spectral densities. It is easy to construct processes that have the desired distance ρ^*. If W_n is an i.i.d. sequence of zero mean, unit variance, Gaussian random variables, then $[A, \mu_\theta]$ can be generated by passing W_n through a linear time-invariant filter with transfer function $f_\theta(\lambda)^{\frac{1}{2}}$. The process $[A, \mu_\varphi]$ can be generated in a similar manner. If we now use W_n to simultaneously generate both X_n and Y_n as shown in the figure, then $Z_n = X_n - Y_n$ can be considered as the output of a linear filter with transfer function $f_\theta(\lambda)^{\frac{1}{2}} - f_\varphi(\lambda)^{\frac{1}{2}}$ and application of linear systems relations (Rozanov (1967) pages 34–35) yields

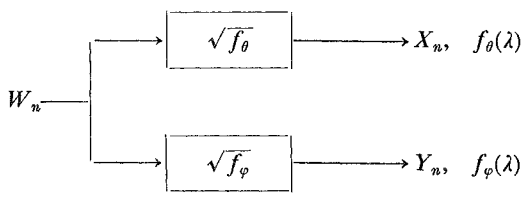

FIG. 1.

that

$$\sigma_Z^2 = E\{(X_n - Y_n)^2\} = (2\pi)^{-1} \int_{-\pi}^{\pi} |f_\theta(\lambda)^{\frac{1}{2}} - f_\varphi(\lambda)^{\frac{1}{2}}|^2 \, d\lambda \, .$$

Since $X_n - Y_n$ is therefore Gaussian with the desired variance, $E((X_n - Y_n)^2) = \bar{\rho}^*(\theta, \varphi)$ and the joint process achieves $\bar{\rho}^*$.

The $\bar{\rho}$ distance may provide a useful topology on Gaussian random processes in problems of system identification and modeling since the equivalent definitions of $\bar{\rho}$ provide interpretations of closeness or goodness of fit in terms of representative sample functions and time average distortion in addition to the ensemble meaning.

We close this section with a corollary:

COROLLARY 3. *Let* $[A, \mu_\theta]$ *and* $[A, \mu_\varphi]$ *be two stationary random processes having power spectral density* $f_\theta(\lambda)$ *and* $f_\varphi(\lambda)$, *then for* $\rho^*(x^1, y^1) = (x^1 - y^1)^2$

$$\bar{\rho}^*(\theta, \varphi) \geqq (2\pi)^{-1} \int_{-\pi}^{\pi} |f_\theta(\lambda)^{\frac{1}{2}} - f_\varphi(\lambda)^{\frac{1}{2}}|^2 \, d\lambda$$

with equality if the processes are Gaussian.

The corollary follows immediately from noticing that the lower bound part of the theorem proof did not involve the fact that the processes were Gaussian. Thus Gaussian processes provide an extreme in the class of stationary processes with given spectra and a mean-squared error "distance" in that we can fit them together at least as well and possibly better than non-Gaussian processes having the same spectra. We conjecture that the lower bound equals $\bar{\rho}^*$ only if the processes are Gaussian.

APPENDIX

PROOF OF THEOREM 1.

(a) $\bar{\rho} = \bar{\rho}''$. Given $\varepsilon > 0$, let $p \in \mathscr{E}_{\theta,\varphi}$ be a measure yielding $E_p[\rho(X_0, Y_0)] \leqq \bar{\rho}''(\theta, \varphi) + \varepsilon$. The restriction p^n of p is contained in $\mathscr{P}_n$ and therefore $E_{p^n}[\rho_n(X^n, Y^n)] \geqq \bar{\rho}_n(\theta, \varphi)$. Since p is assumed stationary, $\bar{\rho}''(\theta, \varphi) + \varepsilon \geqq E_p[\rho(X_0, Y_0)] = E_{p^n}[\rho_n(X^n, Y^n)] \geqq \rho_n(\theta, \varphi)$, all n. Since ε was arbitrary, $\bar{\rho}'' \geqq \bar{\rho}$.

Let $p^n \in \mathscr{P}_n$, $n = 1, 2, \cdots$ be a sequence of measures such that

$$E_{p^n}[\rho_n(X^n, Y^n)] \leqq \bar{\rho}_n + \varepsilon_n \tag{A1}$$

where $\varepsilon_n > 0$ and $\lim_{n\to\infty} \varepsilon_n = 0$. Let q_n denote the product measure on $(\Sigma, \mathscr{S})^2$ induced by the p^n, i.e., for any N and any N-dimensional cylinder $C = \bigtimes_{i=-\infty}^{\infty} C_i$, where all of the $C_i = A^2$ except for a finite number N which are elements of $\mathscr{S}^2$, define

$$q_n(C) = \prod_{j=-\infty}^{\infty} p^n(C_{jn} \times C_{jn+1} \times \cdots \times C_{jn+n-1}) \, .$$

Thus q_n treats successive n-tuples as independent. The T^n-invariant measure q_n can be averaged to form a T-invariant measure p_n on cylinders:

$$p_n(C) = n^{-1} \sum_{i=0}^{n-1} q_n(T^i C) = n^{-1} \sum_{i=0}^{n-1} \prod_{j=-\infty}^{\infty} p^n(C_{jn+i} \times \cdots \times C_{jn+i+n-1}) \, .$$

Since p_n is T-invariant on cylinders and consistent, it extends to a stationary measure, also denoted p_n, on $(\Sigma, \mathscr{S})^2$.

For any $m \leqq n$ the marginal mth restrictions of p_n can be related to the original measure as follows: Let $G = \bigtimes_{k=0}^{m-1} G_k \in \mathscr{G}^m$,

$$q_n(\omega\,;\ \omega = (x, y) \in \Sigma^2;\ x^m \in A^m,\ y^m \in G)$$
$$= p^n(A^n \times (A^{n-m} \times G)) = \mu_\varphi{}^n(A^{n-m} \times G) = \mu_\varphi{}^m(G)$$

since $p^n \in \mathscr{P}_n$. Thus

$$p_n{}^m(A^m \times G) = p_n(\omega :\ \omega = (x, y);\ x^m \in A^m,\ y^m \in G)$$
$$= n^{-1}(n - m + 1)\mu_\varphi{}^m(G) \tag{A2}$$
$$+ n^{-1} \textstyle\sum_{i=1}^{m-1} \mu_\varphi{}^{m-i}(\bigtimes_{k=i}^{m-1} G_k)\mu_\varphi{}^i(\bigtimes_{k=0}^{i-1} G_k)$$

with a similar expression for $G \times A^m$. By a standard diagonalization, the sequence of probability measures p_n has a subsequence p_{n_k} converging on every cylinder in $\mathscr{C}^2$. The limit measure, denoted p, can be extended to a measure on $(\Sigma, \mathscr{S})^2$. From (A2) we have for each fixed m that

$$\lim_{n\to\infty} p_n{}^m(A^m \times G) = p^m(A^m \times G) = \mu_\varphi{}^m(G)$$
$$\lim_{n\to\infty} p_n{}^m(G \times A^m) = p^m(G \times A^m) = \mu_\theta{}^m(G)$$

and hence for any cylinder C in $\mathscr{C}$

$$p(\Sigma \times C) = \mu_\varphi(C)$$
$$p(C \times \Sigma) = \mu_\theta(C)\,.$$

This implies that p induces the desired marginals and hence is itself a probability measure on $(\Sigma, \mathscr{S})^2$. Since by construction p is T-invariant, $p \in \mathscr{E}_{\theta,\varphi}$. Furthermore,

$$\bar{\rho}'' \leqq E_p[\rho(X_0, Y_0)] = \lim_{k\to\infty} E_{p_{n_k}}[\rho(X_0, Y_0)]$$
$$= \lim_{k\to\infty} n_k{}^{-1} \textstyle\sum_{i=0}^{n_k-1} E_{q_{n_k}}[\rho(X_i, Y_i)]$$
$$\leqq \lim_{k\to\infty} (\bar{\rho}_{n_k} + \varepsilon_{n_k}) = \bar{\rho}$$

and therefore $\bar{\rho} \geqq \bar{\rho}''$, completing the proof that $\bar{\rho} = \bar{\rho}''$.

(b) $\bar{\rho}' = \bar{\rho}$. Assume that x is θ-representative and y is φ-representative. Fix n. For each N define for B^n, $C^n \in \mathscr{G}^n$,

$$\mu_N(B^n \times C^n) = N^{-1} \textstyle\sum_{i=0}^{N-1} I_{B^n \times C^n}(x_i, x_{i+1}, \cdots, x_{i+n-1}; y_i, \cdots, y_{i+n-1})\,.$$

Since such rectangles generated $\mathscr{A}^{2n}$, μ_N extends to a measure on $(A^n, \mathscr{A}^n)^2$ and hence for B^n, $C^n \in \mathscr{G}^n$,

$$\lim_{N\to\infty} \mu_N(B^n \times A^n) = \mu_\theta{}^n(B^n)\,, \tag{A3}$$
$$\lim_{N\to\infty} \mu_N(A^n \times C^n) = \mu_\varphi{}^n(C^n)\,.$$

since x and y are representatives.

From Parthasarathy (1968), Theorem 3.2, there is a compact set K such that

$$\mu_\theta{}^n(K) > 1 - \varepsilon\,, \qquad \mu_\varphi{}^n(K) > 1 - \varepsilon\,.$$

Thus (A3) implies that if N is large then $\mu_N(K \times K) > 1 - 4\varepsilon$.

From Theorem 6.7 of Parthasarathy (1968), the sequence $\{\mu_N\}$ therefore has a convergent subsequence of probability measures. Let μ be a limit point of $\{\mu_N\}$. Equation (A3) implies that for B^n, $C^n \in \mathscr{A}^n$,

$$\mu(B^n \times A^n) = \mu_\theta{}^n(B^n),$$
$$\mu(A^n \times C^n) = \mu_\varphi{}^n(C^n),$$

so that $\mu \in \mathscr{P}_n$. It is easy to see that

$$E_{\mu_N}[\rho_n(X^n, Y^n)] = N^{-1} \sum_{j=0}^{N-1} n^{-1} \sum_{i=0}^{n-1} \rho(x_{i+j}, y_{i+j})$$

so that

$$E_\mu[\rho_n(X^n, Y^n)] \leqq \limsup_{N\to\infty} N^{-1} \sum_{i=0}^{N-1} \rho(x_i, y_i)$$

and hence $\bar{\rho} \leqq \bar{\rho}'$.

Choose for $\varepsilon > 0$ a stationary measure $p \in \mathscr{E}_{\theta,\varphi}$ such that

$$E_p(\rho(X_0, Y_0)) \leqq \bar{\rho}'' + \varepsilon$$

since p is stationary and has marginals μ_θ and μ_φ relative frequencies converge, for almost all points $\omega = (x, y) \in \Sigma^2$ such that x is θ-representative, y is φ-representative, and

$$\hat{f}(\omega) = \lim_{n\to\infty} n^{-1} \sum_{i=0}^{n-1} \rho(x_i, y_i)$$

exists and therefore

$$\hat{f}(\omega) \geqq \bar{\rho}'.$$

Taking expectation under p yields from the ergodic theorem and the above that

$$\bar{\rho}' \leqq E_p[\hat{f}(\omega)] = E_p[\rho(x_0, y_0)] \leqq \bar{\rho}'' + \varepsilon.$$

Since ε is arbitrary $\bar{\rho} = \bar{\rho}'' \geqq \bar{\rho}'$, completing the proof.

Acknowledgment. The authors wish to thank Professors Lee Davisson of U.S.C., and Don Ornstein and Thomas Kailath of Stanford University for many helpful suggestions and comments.

REFERENCES

BERGER, T. (1971). *Rate Distortion Theory*. Prentice-Hall, Englewood Cliffs.

BLUMENTHAL, R. M. (1973). Discussion on Professor Ornstein's paper. *Ann. Probability* **1** 60.

DOBRUSHIN, R. L. (1963). Unified methods for transmission of information; the general case. *Soviet Math.* **4** 289–292.

GALLAGER, R. G. (1968). *Information Theory and Reliable Communication*. Wiley, New York.

GRAY, R. M. and DAVISSON, L. D. (1974). Source coding without the ergodic assumption. *IEEE Trans. Information Theory* **20** 502–516.

GRAY, R. M., NEUHOFF, D. L. and OMURA, J. K. (1975). Process definitions of the distortion-rate function and the source coding theorem. To appear in *IEEE Trans. Information Theory* **21**.

GRAY, R. M., NEUHOFF, D. L. and ORNSTEIN, D. S. (1975). Non-block source coding with a fidelity criterion. *Ann. Probability* **3** No. 3.

KAILATH, T. (1973). Private communication.

KRENGEL, U. (1973). Discussion of Professor Ornstein's paper. *Ann. Probability* **1** 61–62.

NEUHOFF, D. L., GRAY, R. M. and DAVISSON, L. D. (1975). Fixed rate universal source coding with a fidelity criterion. To appear in *IEEE Trans. Information Theory* **21**.

ORNSTEIN, D. S. (1973). An application of ergodic theory to probability theory. *Ann. Probability* **1** 43–58.

PARTHASARATHY, K. R. (1968). *Probability Measures on Metric Spaces.* Academic Press, New York, Chapter 3.

PINSKER, M. S. (1960). *Information and information stability of random variables. Izd. Akad. Nauk, SSSR, Moscow.*

ROZANOV, Y. (1967). *Stationary Random Processes.* Holden-Day, San Francisco.

SAKRISON, D. J. (1969). The rate distortion function of a class of sources. *Information and Control* **15** 165–195.

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

SHANNON, C. E. (1959). Coding theorems for u discrete source with a fidelity criterion. *IRE National Convention Record, Part* 4. 142–163.

SHIELDS, P. (1974). Separability of the space of Markov chains in the $\bar{d}$ metric. Submitted for publication.

SHIELDS, P. and ENGLAND, J. (1974). Paper in preparation.

VALLENDER, S. S. (1973). Computing the Wasserstein distance between probability distributions on the line. *Theor. Probability Appl.* **18** 824–827 (in Russian).

VASERSHTEIN, L. N. (1969). Markov processes on countable product space describing large systems of automata. *Problemy Peredachi Informatsii* **5** 64–73.

ZIV, J. (1972). Coding of sources with unknown statistics II: distortion relative to a fidelity criterion. *IEEE Trans. Information Theory* **18** 389–394.

ROBERT M. GRAY
DEPARTMENT OF ELECTRICAL ENGINEERING
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

PAUL C. SHIELDS
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF TOLEDO
TOLEDO, OHIO

DAVID L. NEUHOFF
DEPT. OF ELECTRICAL ENGINEERING
AND COMPUTER SCIENCE
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN 48104