Rate Distortion Theory for Sources with Abstract Alphabets and Memory

Toby Berger†

Advanced Development Laboratory, Raytheon Company
Wayland, Massachusetts 01778

This paper is devoted to the formulation and proof of an abstract alphabet version of the fundamental theorem of Shannon's rate distortion theory. The validity of the theorem is established for both discrete and continuous parameter information sources satisfying a certain regularity condition intermediate in restrictiveness between ergodicity and weak mixing. For ease of presentation, only single letter fidelity criteria are considered during the main development after which various generalizations are indicated.

I. INTRODUCTION

The rate distortion function was introduced by Shannon (1948, 1959) to specify the rate at which a source produces information relative to a fidelity criterion. Rate distortion theory is an essential tool in the study of analog data sources and, more generally, is applicable to any communication problem in which the entropy rate of the source exceeds the capacity of the channel over which it must be transmitted. Since perfect transmission is impossible in such situations, it becomes fruitful to study the problem of approximating the given source with one of lower entropy in such a way that the least possible distortion results relative to some prescribed fidelity criterion.

For the case of a bounded distortion measure and a discrete parameter source producing statistically independent samples, Shannon (1959) defined the rate distortion function, \( R(\cdot) \), and then proved that \( R(D) \) represents the minimum capacity a channel must have in order for it to be possible to reproduce the source at the channel output with an average distortion no greater than \( D \). The extension to discrete parameter sources with memory was briefly sketched also, though no consideration

† Present address: School of Electrical Engineering, Cornell University, Ithaca, New York 14850.
was given to the statistical dependence between successive blocks of source letters used in the random coding argument.

In the present paper we provide a detailed proof of an abstract alphabet formulation of the fundamental theorem of rate distortion theory for discrete or continuous parameter sources and unbounded distortion measures. The statistical dependence between successive blocks of letters produced by the source necessitates our imposing a regularity condition on the source called block ergodicity (see Definition 1), which is intermediate in restrictiveness between ordinary ergodicity and weak mixing. Our main result (Theorem 2) may be interpreted as formally establishing that the source-fidelity criterion combinations in question possess the property of information stability required for application of the basic theorems presented by Dobrushin (1959) in his generalized formulation of information theory.

While this paper was undergoing review Goblick (1967) reported similar results for time discrete strongly mixing sources. Still more recently Gallager (1968) has successfully treated time discrete finite alphabet ergodic sources by means of a decomposition technique devised by Nedoma (1963).

The terminology and notation employed are presented in Section II together with certain preliminary results. The rate distortion function of an information source with respect to a (single letter) fidelity criterion is defined in Section III. The fundamental theorem of rate distortion theory, which imbues the rate distortion function with its operational significance, is stated and proved in Section IV. Some straightforward generalizations of the results are indicated in the final section.

II. TERMINOLOGY, NOTATION, AND PRELIMINARIES

Let the space $Z_0$, called the joint alphabet, be the Cartesian product of two non-empty abstract spaces $X_0$ and $Y_0$ called the message alphabet and the approximating alphabet, respectively. Let $Z_0$ be the product $\sigma$-algebra of subsets of $Z_0$ derived from the $\sigma$-algebras $\mathcal{X}_0$ and $\mathcal{Y}_0$ of subsets of $X_0$ and $Y_0$, respectively. Let the measurable space $(Z, Z) = \prod_{t \in \mathcal{M}} (Z_t, Z_t)$ be the infinite Cartesian product of exemplars $(Z_t, Z_t)$ of the measurable space $(Z_0, Z_0)$, where the index set $\mathcal{M}$ is either the integers (discrete time) or the real line (continuous time). Similarly, let $(X, \mathcal{X}) = \prod_{t \in \mathcal{M}} (X_t, \mathcal{X}_t)$
and

\[(Y, y) = \prod_{t \in M} (Y_t, y_t),\]

from which it follows that \((Z, z) = (X \times Y, \mathcal{X} \times \mathcal{Y})\).

Let \(\mu(\cdot)\) be a probability on the \(\sigma\)-algebra \(\mathcal{X}\), that is, a nonnegative, countably additive set function such that \(\mu(X) = 1\). We call the triple \([X, \mathcal{X}, \mu]\) the message source, or simply the source whenever no ambiguity results. Following McMillan (1953), we abbreviate the symbol for the source to \([X, \mu]\). Elements \(x \in X\) are called realizations of the source. We use the symbol \(\pi[\mu]\) to indicate that the proposition \(\pi(x)\) is true for almost every \(x\) with respect to \(\mu\).

The value assumed by the projection of a realization \(x \in X\) into the \(t\)th coordinate space is denoted by \(x_t \in X_t\). Similarly, the \(t\)th coordinates of \(y \in Y\) and \(z = (x, y) \in Z\) are denoted by \(y_t\) and \(z_t\), respectively.

For any \(r, t \in M\), with \(r < t\), we define the probability space \((X_{r,t}, \mathcal{X}_{r,t}, \mu_{r,t})\) by

\[(X_{r,t}, \mathcal{X}_{r,t}) = \prod_{s \in [r,t)} (X_s, \mathcal{X}_s)\]

with \(\mu_{r,t}\) being the restriction of \(\mu\) to \(\mathcal{X}_{r,t}\). The measurable spaces \((Y_{r,t}, \mathcal{Y}_{r,t})\) and \((Z_{r,t}, \mathcal{Z}_{r,t})\) are defined by expressions analogous to (1). For completeness, we set \((X_{t,t}, \mathcal{X}_{t,t}) = (X_t, \mathcal{X}_t)\), with analogous notation for the approximating and joint spaces, too. For the particular case of \(r = 0\) a single, superscript \(t\) is used, e.g.,

\((X_{0,t}, \mathcal{X}_{0,t}, \mu_{0,t}) = (X^t, \mathcal{X}^t, \mu^t)\).

Typical elements of \(X^t[Y^t, Z^t]\) are called message [approximating, joint] \(t\)-blocks and are denoted by \(x^t[y^t, z^t = (x^t, y^t)]\).

We indulge in the following abuses of notation. The symbol \(x_t\) used to denote the \(s\)th component of a realization \(x \in X\) also is used to denote the \(s\)th component of a typical \(t\)-block, \(x^t, 0 \leq s < t\). Moreover, the superscript \(t\) of \(x^t, y^t\) and \(z^t\) often is suppressed, since one or two judiciously placed \(t\)'s generally suffice to permit relatively lengthy expressions to be interpreted unambiguously.

Let \(\{T^t, t \in M\}\) be the group of shift transformations from \(X\) to \(X\) defined by \((T^tx)_s = x_{s+t}\).\(^1\) A source \([X, \mu]\) is called \(t\)-stationary if \(\mu(T^tE) = \mu(E)\) for all \(E \in \mathcal{X}\), where \(T^tE = \{T^tx: x \in E\}\). We say \([X, \mu]\)

\(^1\) The symbol \(T^t\) also will be used to denote the shift by \(t\) coordinates on the spaces \(Y\) and \(Z\).
is stationary if it is $t$-stationary for all $t \in M$. A $\mathcal{X}$-measurable function $g$ is called $t$-invariant if $g(T^t x) = g(x)$ for all $x \in X$, and invariant if it is $t$-invariant for every $t \in M$. A set $E \in \mathcal{X}$ is $t$-invariant (invariant) if its characteristic function is $t$-invariant (invariant). A source $[X, \mu]$ is $t$-ergodic (ergodic) if $\mu(E) = \mu^2(E)$ for every $t$-invariant (invariant) set $E$.

We shall subsequently be concerned with encoding the source $[X, \mu]$ by means of operations on successive message blocks. In this regard consider segmenting each realization $x \in X$ into an infinite sequence of $\tau$-blocks. The space of all such infinite sequences of $\tau$-blocks, together with the $\sigma$-algebra and probability naturally induced by $\mathcal{X}$ and $\mu$, will be denoted by $[X, \mu]$, and called the $\tau$-block source derived from $[X, \mu]$. The source $[X, \mu]$ is time discrete, producing one message $\tau$-block every $\tau$ seconds.

**Definition 1.** A source $[X, \mu]$ is block ergodic if it is $\tau$-ergodic for every positive $\tau \in M$.

We show in the appendix that block ergodicity lies between ergodicity and weak mixing in restrictiveness. Note that $\tau$-ergodicity of $[X, \mu]$ and ergodicity of $[X, \mu]$, are equivalent.

Let the mapping $q^t : X^t \times Y^t \rightarrow [0, 1]$ be a transition probability, by which we mean that $q^t$ satisfies the following two conditions:

(a) For every $x \in X^t$, the set function $q^t(x, \cdot)$ is a probability on $Y^t$.

(b) For every $F \in Y^t$, the function $q^t(\cdot, F)$ is $\mathcal{X}^t$-measurable.

Let $\omega^t$ denote the joint probability induced on $Z^t$ by $\mu^t$ and $q^t$. That is, for any set $G \in Z^t$, we have:

$$\omega^t(G) = \int q^t(x, G_x) \, d\mu^t,$$

where $G_x = \{y : (x, y) \in G\}$. We denote the marginal probability that $\omega^t$ induces on $Y^t$ by

$$\nu^t(F) = \omega^t(X^t \times F) = \int q^t(x, F) \, d\mu^t,$$

and the product probability on $Z^t$ derived from the marginals of $\omega^t$ by

$$w^t = \mu^t \times \nu^t.$$

If $\omega^t$ is absolutely continuous with respect to $w^t$ (henceforth written

\[2\] Unless subscripted by a particular set, integrals with respect to a probability extend over the entire space governed by that probability.
\( \omega^t \ll w^t \), then we denote the associated Radon–Nikodym derivative by

\[
\Pi_t = d\omega^t / dw^t.
\]

If \( \omega^t \ll w^t \), then for any \( G \in Z^t \) we have

\[
\omega^t(G) = \int_G \Pi_t \, dw^t = \int_G \Pi_t \, d(\mu^t \times \nu^t) = \int \left( \int_{G_x} f_t(x, y) \, dw^t \right) \, d\mu^t.
\]

(3)

Comparing (2) and (3), we note for future reference that

\[
q^t(x, G_x) = \int_{G_x} f_t(x, y) \, dw^t [\mu^t].
\]

(4)

DEFINITION 2. The average information, \( I_t \), of the joint probability space \((Z^t, Z^t, \omega^t)\) is defined by

\[
I_t = \sup \sum_{i=1}^{\infty} \omega^t(G_i) \log \left[ \omega^t(G_i)/w^t(G_i) \right],
\]

(5)

where the supremum is taken with respect to all partitions \( \{G_i\} \) of the space \( Z^t \) by countably many rectangles \( G_i = E_i \times F_i \) with \( E_i \subseteq \mathfrak{X}^t \) and \( F_i \subseteq \mathfrak{Y}^t \). Dobrushin (1959) shows that the value of \( I_t \) always is non-negative and remains unchanged if the restriction to rectangular partitions is relaxed. When \( \omega^t \) and \( w^t \) are induced by a transition probability \( q^t \), then \( I_t \) becomes a functional of \( q^t \) which we shall denote by \( I_t(q) \).

LEMMA 1. If \( I_t \) is finite, then \( \omega^t \ll w^t \) and

\[
I_t = \int \log \Pi_t \, d\omega^t.
\]

(6)

Proof. See the translator's remarks by Feinstein in the book by Pinsker (1960).

We shall need the so-called generalized Shannon–McMillan limit theorem and its corollary, the generalized asymptotic equipartition property (AEP). The following version of the theorem, due to Perez (1964), is essentially the most general to be established to date.

THEOREM 1. (Perez's Theorem). Let the joint source \([Z, Z, \omega]\) be stationary, and let \((Z^t, Z^t, \omega^t)\) be its restriction to joint \( t \)-blocks. Then the finiteness of the information rate,

\[
R \triangleq \lim_{t \to \infty} (1/t) I_t,
\]

(7)

1 All logarithms in this paper are to the base \( e \).
is a necessary and sufficient condition for the existence of an invariant, ω-integrable function $h(z)$ such that

$$\lim_{t \to \infty} (1/t) \log f_t(z) = h(z) [\omega]. \quad (8)$$

If $[Z, Z, \omega]$ is ergodic as well, then $h(z)$ is a constant, namely the information rate $R$ of (7).

**Corollary 1. (Generalized AEP).** If the joint source $[Z, Z, \omega]$ of Theorem 1 also is ergodic and $R$ of (7) is finite, then for all $\epsilon > 0$

$$\lim_{t \to \infty} \omega \{z: \mid (1/t) \log f_t(z) - R \mid > \epsilon \} = 0. \quad (9)$$

### III. The Rate Distortion Function

In this section we define the rate distortion function of an information source with respect to a single letter fidelity criterion.

Any $Z_0$-measurable function $\rho$ mapping $Z_0$ into $[0, \infty)$ is called a single letter distortion measure. Its physical significance is that $\rho(\alpha, \beta)$ specifies the distortion that results when $\alpha \in X_0$ is approximated by $\beta \in Y_0$. We associate with $\rho$ a family of distortion measures,

$$F_\rho = \{\rho_t, 0 \leq t < \infty\}, \quad (10)$$

each member of which maps $Z$ into $[0, \infty)$. Specifically, we set $\rho_0(z) = \rho(z_0)$, and for $t > 0$ we define

$$\rho_t(z) = \frac{1}{t} \sum_{s=0}^{t-1} \rho_0(T^s z) \quad \text{or} \quad \frac{1}{t} \int_0^t \rho_0(T^s z) \, ds \quad (11)$$

for discrete or continuous time, respectively. Since $\rho_t(z)$ depends only on those components of $z$ that belong to $z^t$, it is also possible to interpret it as a $Z^t$-measurable mapping from $Z^t$ into $[0, \infty)$ and we often shall.

**Definition 3.** The family of distortion measures $F_\rho$ of (10) is called a single letter fidelity criterion.

We proceed to define the rate distortion function of a message source with respect to a single letter fidelity criterion. Toward this end let us introduce the family of functions, $\{R_t(D), 0 \leq t < \infty\}$, defined as follows:

For each $D \in [0, \infty)$, let $Q_t(D)$ be the class of transition probabilities $q^t: X^t \times Y^t \to [0, 1]$ for which

4 The symbol "$f_t$" in (8) denotes a function defined on $Z$, not $Z^t$. By rights this function should be given a different symbol, say $f_t$, and then defined by $f_t(z) = f_t(z^t)$, but it is customary to use the same symbol for both functions.
That is, \( q^t \) belongs to \( Q_t(D) \) only if it induces a joint probability \( \omega^t \) on \( Z_t \) for which the average distortion \( D(q) \) between message \( t \)-blocks and approximating \( t \)-blocks does not exceed \( D \). We also introduce the average information rate \( R_t(q) \) defined, with reference to (5), to be

\[
R_t(q) = \frac{1}{t} \left[ I_t(q) \right],
\]

where the value \( +\infty \) is not excluded. The function \( R_t(D) \) then is defined by

\[
R_t(D) = \inf_{Q_t(D)} R_t(q).
\]

If \( Q_t(D) \) is empty, \( R_t(D) \) is taken to be \( +\infty \).

**Definition 4.** The rate distortion function \( R(D) \) of the message source \( [X, \mu] \) with respect to the single letter fidelity \( F_p \) is given by the prescription

\[
R(D) = \liminf_{t \to \infty} R_t(D).
\]

For completeness we record here the following facts concerning \( R(D) \) (Shannon, 1959). First, the "inf" may be deleted in (15) provided \( [X, \mu] \) is stationary (see also Reiffen, 1966). Second, \( R(D) \) is a nonnegative, monotonic nonincreasing, convex downward function of \( D \) over the interval in which it is finite. Third, if there is an element \( y \in Y \) for which

\[
\lim_{t \to \infty} \int \rho_t(x, y) d\mu < \infty,
\]

then there is a distortion value \( D_{\text{max}} \) such that \( R(D) \) vanishes identically for \( D \geq D_{\text{max}} \).

It is heuristically clear from Definition 4 that \( R(D) \), if finite, in some sense represents the minimum average information per unit time that must be supplied about a realization \( x \in X \) in order to permit specification of a \( y \in Y \) that approximates this \( x \) with an average distortion with respect to \( F_p \) that does not exceed \( D \). A precise mathematical statement of the sense in which this heuristic interpretation of \( R(D) \) is correct is provided by the fundamental theorem of rate distortion theory.

**IV. THE FUNDAMENTAL THEOREM OF RATE DISTORTION THEORY**

In this section we prove our main result (Theorem 2) and then discuss some of its implications.
THEOREM 2. If there exists \( \alpha \in Y_0 \) such that
\[
\int \rho_0(x, \beta) \, d\mu = \beta < \infty,
\] (16)
then the following two statements are valid for any \( \epsilon > 0 \) and any \( D \geq 0 \):

Positive Statement. If \([X, \mu]\) is both stationary and block ergodic and \( R(D) \) is finite, then there exists a value of \( t \) and a subset \( S \) of \( Y^t \) containing \( N \) approximating \( t \)-blocks such that \( \log N \leq t[R(D) + \epsilon] \) and
\[
D(S) \triangleq \int \rho_t(x | S) \, d\mu^t \leq D + \epsilon,
\] (17)
where
\[
\rho_t(x | S) = \min_{y \in S} \rho_t(x, y).^5
\] (18)

Negative Statement. For all \( t \) any set \( \hat{S} \subset Y^t \) that contains only \( \hat{N} \) elements, where \( \log \hat{N} \leq t[R(D) - \epsilon] \), must satisfy the inequality \( D(\hat{S}) > D \).

Proof. We shall prove the positive statement by a random coding argument. Consider an ensemble of subsets \( S \) of \( Y^t \) each of which contains \( N \) elements, where \( N \) is the largest integer in \( \exp \{t[R(D) + \epsilon] \} \). If we can show that the average over this ensemble of \( D(S) \) as defined in (17) does not exceed \( D + \epsilon \), then at least one \( S \) in the ensemble must satisfy inequality (17) and the proof will be complete.

We define our ensemble of subsets of \( Y^t \) as follows. The choice of a particular set of \( N \) elements of \( Y^t \) may be envisioned as the choice of a single element \( S \in Y^t \). We choose each such \( S \) independently of all the others according to a common probability measure \( \lambda^{tN} \) defined on \( Y^{tN} \). Furthermore we select the \( N \) elements of each \( S \) independently, i.e.,
\[
\lambda^{tN} = \nu_1^t \times \nu_2^t \times \cdots \times \nu_N^t,
\] (19)
where each \( \nu_i^t \) is a probability on \( Y^t \). Moreover, for \( i = 1, 2, \cdots, N - 1 \), we set each \( \nu_i^t \) equal to a common probability, \( \nu^t \). The probability \( \nu_N^t \) that governs the selection of the last element of each set is a degenerate one concentrated at a special element \( b \in Y^t \), where \( b \) may be taken as any approximating \( t \)-block for which
\[
\int \rho_t(x, b) \, d\mu^t < \infty.
\] (20)

In particular, one candidate for \( b \) is that element every component of

---

^ Either, but not both, of the \( \epsilon \)'s in the positive statement may be set to zero by a straightforward extension of the proof given below.
which is the \( \beta \) that appears in (16), in which case the integral in (20) equals \( \beta \). Since this choice of \( b \) simplifies subsequent arguments, we set \( b_s = \beta, 0 \leq s < t \). Recalling (17) and (18), we see that the average of \( D(S) \) over the ensemble of sets \( S \) may be written

\[
\bar{D} = \int D(S) \, d\lambda^t = \int \left( \int \rho_t(x \mid S) \, d\mu^t \right) \, d\lambda^t.
\]

We now define the set

\[
A = \{ x \in X^t : \rho_t(x, b) \leq \beta + \delta \},
\]

where \( \delta \) is an arbitrary positive constant. The fact that \( b \) belongs to every \( S \) in the ensemble guarantees that

\[
\rho_t(x \mid S) \leq \rho_t(x, b).
\]

Accordingly, where \( \bar{A} \) denotes the complement of \( A \), we have

\[
\bar{D} \leq \int_{\bar{A}} \rho_t(x, b) \, d\mu^t + \int \left( \int_{\bar{A}} \rho_t(x \mid S) \, d\mu^t \right) \, d\lambda^t.
\]

Now (20) and (22) imply that the iterated integral appearing on the far right of (23) is finite, so Fubini's theorem yields

\[
\int \left( \int_{\bar{A}} \rho_t(x \mid S) \, d\mu^t \right) \, d\lambda^t = \int \left( \int_{\bar{A}} \rho_t(x \mid S) \, d\lambda^t \right) \, d\mu^t.
\]

The inner integral on the right of (24) is the ensemble average of \( \rho_t(x \mid S) \) for fixed \( x \). Hence, letting

\[
P_t = P_t(u \mid x) = \lambda^t\{S : \rho_t(x \mid S) \leq u\}
\]

denote the cumulative probability distribution function of \( \rho_t(x \mid S) \) for fixed \( x \), we may write

\[
\int \left( \int_{\bar{A}} \rho_t(x \mid S) \, d\mu^t \right) \, d\lambda^t = \int \left( \int_0^{\hat{\beta} + \delta} u \, dP_t \right) \, d\mu^t,
\]

where the finite upper limit of \( \hat{\beta} + \delta \) follows from (22) and the definitions of \( A \) and \( P_t \). Upon introducing the complementary distribution function

\[
\bar{P}_t = 1 - P_t,
\]

we note that

\[
\int_0^{\hat{\beta} + \delta} u \, dP_t = - \int_0^{\hat{\beta} + \delta} u \, d\bar{P}_t = - (\hat{\beta} + \delta)\bar{P}_t(\hat{\beta} + \delta \mid x) + \int_0^{\hat{\beta} + \delta} \bar{P}_t \, du,
\]
the last integral being with respect to ordinary Lebesgue measure. Thus,
\[
\int_{0}^{\beta + \delta} u \, dP_t \leq \int_{0}^{\beta + \delta} \tilde{P}_t \, du \leq (D + \delta) + \int_{D + \delta}^{\beta + \delta} \tilde{P}_t \, du \leq D + \delta + (\hat{\rho} - D)\tilde{P}_t(D + \delta \mid x),
\]  
where we have used the fact that \( \tilde{P}_t(u \mid x) \) is a monotonic nonincreasing function of \( u \) that never exceeds unity. (Of course, we are also assuming that \( \hat{\rho} > D \), since the theorem is trivial for \( \hat{\rho} \leq D \).) Substituting (26) into (25) and upper bounding the integration over \( X \) allows us to replace (23) by
\[
\bar{D} \leq D + \delta + \int_{A} \rho_t(x, b) \, d\mu^t + (\hat{\rho} - D) \int \tilde{P}_t(D + \delta \mid x) \, d\mu^t. \tag{27}
\]

We complete the proof by showing that both integrals in (27) vanish in the limit of infinite \( t \). The integral over \( \bar{A} \) may be dispensed with by means of the ergodic theorem. Toward this end we write
\[
\int_{\bar{A}} \rho_t(x, b) \, d\mu^t = \int_{\bar{A}} [\rho_t(x, b) - \hat{\rho}] \, d\mu^t + \hat{\rho} \mu^t(\bar{A})
\]
\[
= \int [\rho_t(x, b) - \hat{\rho}] \, d\mu^t - \int_{A} [\rho_t(x, b) - \hat{\rho}] \, d\mu^t + \hat{\rho} \mu^t(\bar{A})
\]
\[
= \int_{A} [\hat{\rho} - \rho_t(x, b)] \, d\mu^t + \hat{\rho} \mu^t(\bar{A}),
\]  
where we have used the fact that the expected value of \( \rho_t(x, b) \) is \( \hat{\rho} \) for our choice of \( b \). Let
\[
B = \{ x \in X^t : \rho_t(x, b) \geq \hat{\rho} - \delta \},
\]
and observe that
\[
\int_{A} [\hat{\rho} - \rho_t(x, b)] \, d\mu^t = \int_{A \cap B} [\hat{\rho} - \rho_t(x, b)] \, d\mu^t
\]
\[
+ \int_{A \cap \bar{B}} [\hat{\rho} - \rho_t(x, b)] \, d\mu^t \leq \delta \mu^t(A \cap B) + \hat{\rho} \mu^t(A \cap \bar{B}) \leq \delta + \hat{\rho} \mu^t(\bar{B}). \tag{29}
\]
Combining (28) and (29) and noting that $\tilde{A}$ and $\tilde{B}$ are disjoint yields
\[
\int_{\tilde{A}} \rho_t(x, b) \, d\mu^t \leq \delta + \hat{\rho} \mu^t(\tilde{A} \cup \tilde{B}). \tag{30}
\]
Since $\tilde{A} \cup \tilde{B} = \{x : |\rho_t(x, b) - \hat{\rho}| > \delta\}$, it follows in the case of, say, continuous time that
\[
\mu^t(\tilde{A} \cup \tilde{B}) = \mu\left\{x \in X : \left| \frac{1}{t} \int_0^t \rho_0(T^s x, \beta) \, ds - \hat{\rho} \right| > \delta \right\}. \tag{31}
\]
Because $[X, \mu]$ is ergodic, the first term within the absolute value brackets in (31) converges to the expected value of $\rho_0(x, \beta)$, namely $\hat{\rho}$ of (16). Thus, $\mu^t(\tilde{A} \cup \tilde{B}) \to 0$, and the integral in (30) vanishes for large $t$ since $\delta$ is arbitrary.

We now show that the other integral in (27), namely
\[
J_t \triangleq \int \tilde{P}_t(D + \delta \mid x) \, d\mu^t, \tag{32}
\]
also vanishes in the limit of large $t$. We begin by noting that excluding the special element $b$ from consideration when determining the minimum in (18) yields the inequality
\[
\tilde{P}_t(D + \delta \mid x) = \lambda^N \{S : \rho_t(x \mid S) > D + \delta\} \leq [1 - \nu^t(y : \rho_t(x, y) \leq D + \delta)]^{N-1}, \tag{33}
\]
where $\nu^t$ is the common probability governing the independent selection of the other $N - 1$ elements of $S$.

At this point it becomes necessary to discuss in greater detail the manner in which we select $\nu^t$. First, we choose a block length $\tau$ large enough to ensure that
\[
R_\tau(D) < R(D) + \delta/2. \tag{34}
\]
Next, recalling (13) and (14), we choose a transition probability $q^t \in \mathcal{Q}_\tau(D)$ for which
\[
\omega_\tau(q) < R_\tau(D) + \delta/2. \tag{35}
\]
In the ensuing discussion $q^t$ is referred to as the "$\tau$-block channel." As usual, $\omega^t$ and $\nu^t$ denote the probabilities that $q^t$ induces on $\mathbb{Z}$ and $\mathbb{Y}$, respectively.

Since our concern is with the behavior of $J_t$ as $t \to \infty$, no loss in generality results from confining attention solely to values of $t$ of the
form \( t = n\tau, n = 1, 2, \cdots \). For such \( t \) consider the transition probability

\[
q^t = q_1^\tau \times q_2^\tau \times \cdots \times q_n^\tau,
\]

where each \( q_k^\tau \) equals \( q^\tau \). That is, \( q^t \) is the \( n \)-fold product of the \( \tau \)-block channel with itself, and corresponds physically to segmenting each source \( t \)-block into \( n \) successive \( \tau \)-blocks and then transforming these independently with the \( \tau \)-block channel. The probability \( \nu^t \) induced by \( q^t \) on \( \gamma^t \) is the one we use to randomly generate the approximating \( t \)-blocks of our ensemble. Although \( q^t \) operates on successive message \( \tau \)-blocks independently of one another, the source \([X, \mu]\) in general produces message \( \tau \)-blocks that are mutually dependent. As a result, \( \nu^t \) is not the \( n \)-fold product of \( \nu^\tau \) with itself. Since this dependence can only decrease the average information \( I_t(q) \) below the value \( nI_\tau(q) \) it would assume if successive message \( \tau \)-blocks were indeed statistically independent and identically distributed in accordance with \( \mu^\tau \), we have

\[
(1/t)I_t(q) \leq (1/\tau)I_\tau(q) = \Theta_\tau(q) < R(D) + \delta.
\]

We have assumed in the theorem statement that \( R(D) \) is finite, so \((36)\) implies that \( I_t \) is finite for every \( t \). This permits us to conclude from Lemma 1 that \( \omega^t \ll \omega^t \) and, therefore, that the Radon–Nikodym derivative \( f_t = d\omega^t/d\omega^t \) exists and is unique \([\omega^t]\).

We now introduce the sets

\[
\Delta = \{z \in Z^t: \rho_t(z) \leq D + \delta\}
\]

and

\[
\Gamma = \{z \in Z^t: f_t(z) \leq e^{t[R(D)+\delta]}\},
\]

and their respective \( \gamma^t \)-measurable sections \( \Delta_x = \{y: (x, y) \in \Delta\} \) and \( \Gamma_x = \{y: (x, y) \in \Gamma\} \). Returning to \((33)\) we deduce with the aid of \((4)\) that

\[
\nu^t[y: \rho_t(x, y) \leq D + \delta] = \nu^t(\Delta_x) = \int_{\Delta_x} dv^t \geq e^{-t[R(D)+\delta]} \int_{\Delta_x \cap \Gamma_x} f_t(x, y) dv^t = e^{-t[R(D)+\delta]} q^t(x, (\Delta \cap \Gamma)_x),
\]

where we have used that fact that \( \Delta_x \cap \Gamma_x = (\Delta \cap \Gamma)_x \). Inequality \((37)\) may be used to cast \((33)\) in the form

\[
P(D + \delta | x) \leq [1 - e^{-t[R(D)+\delta]} q^t(x, (\Delta \cap \Gamma)_x)]^{N-1}.
\]
At this point we digress momentarily to establish an inequality which we formulate as

**Lemma 2.** If $0 \leq \xi, \gamma \leq 1$ and $K \geq 1$, then

$$p(\xi, \gamma) = (1 - \xi\gamma)^K \leq 1 - \gamma + (1 - \xi)^K.$$  

**Proof.** $p(\xi, \gamma)$ is a convex downward function of $\gamma \in [0, 1]$ for fixed $\xi \in [0, 1]$ because $\partial^2 p / \partial \gamma^2 = (K)(K - 1)\xi^2(1 - \xi\gamma)^{K-2} \geq 0$. Accordingly, $p(\xi, \gamma) \leq (1 - \gamma)p(\xi, 0) + \gamma p(\xi, 1) \leq 1 - \gamma + (1 - \xi)^K$.

Application of Lemma 2 and the well-known inequality $\log x \leq x - 1$ to (38) yields

$$\bar{P}(D + \delta | x) \leq 1 - q^t(x, (\Delta \cap \Gamma)_x) + \exp \{-N - 1)e^{-(R(D) + \delta)}\}. \quad (39)$$

Since $N$ has been chosen as the largest integer in $\exp \{-\epsilon[R(M) + \epsilon]\}$, choosing $\delta < \epsilon$ drives the last term in (39) to zero at a double exponential rate as $t \to \infty$. Hence, if we can show that

$$\lim_{t \to \infty} \int [1 - q^t(x, (\Delta \cap \Gamma)_x) ] \, d\mu^t = 0,$$

then it will follow from (38) and (47) that $J_t \to 0$. Toward this end we observe from (2) that

$$\int [1 - q^t(x, (\Delta \cap \Gamma)_x) ] \, d\mu^t = 1 - \omega^t(\Delta \cap \Gamma) \leq \omega^t(\bar{\Delta}) + \omega^t(\bar{\Gamma}).$$

Thus, it suffices to show that the joint probabilities of the sets $\bar{\Delta}$ and $\bar{\Gamma}$ vanish in the limit of large $t$. In this regard, consider the joint $\tau$-block source $[Z, \omega]$, which results from repeated use of the $\tau$-block channel to transform the successive $\tau$-blocks produced by $[X, \mu]$. Our assumption that $[X, \mu]$ is stationary and block ergodic assures that $[X, \mu]$ is, too. Therefore, $[Z, \omega]$, is both stationary and ergodic because it is the result of a memoryless operation on the time discrete stationary ergodic source, $[X, \mu]$ (Alder, 1961). Since $t = n\tau$, we have

$$\omega^t(\bar{\Delta}) = \omega\{z \in Z: \rho_{n\tau}(z) > D + \delta\} = \omega\left\{z \in Z: \frac{1}{n} \sum_{k=0}^{n-1} \rho_{\tau}(T^{kr}z) > D + \delta\right\}. \quad (40)$$

The ergodic theorem applied to $[Z, \omega]$, implies that in the limit of large $n$ (large $t$) the normalized sum in (40) converges in probability to the
expected value of $\rho_r(z)$, i.e., to the average distortion of a single $r$-block. This, of course, is simply the average distortion $D_r(q)$ associated with the $r$-block channel, $q'$. But $D_r(q) \leq D$ because $q' \in \xi_r(D)$, so $\omega^r(\Delta) \to 0$ as desired.

As regards $\Gamma$ we may write

$$\omega^r(\Gamma) = \omega^r_r\left\{z: \frac{1}{n_r} \log f_{n_r}(z) > R(D) + \delta \right\}. \quad (41)$$

The next step clearly is to apply the AEP. Note, however, that the generalized AEP as given by Corollary 1 is not directly applicable to the joint source $[Z, \omega]$ induced by $[X, \mu]$ and repeated use of the $r$-block channel because $[Z, \omega]$ is not stationary. On the other hand, Corollary 1 does apply to $[Z, \omega_r]$ which, as we have previously noted, is both stationary and ergodic under our assumptions. Hence, with $[Z, \omega_r]$ playing the role of the joint source in Corollary 1, we see that $(n_r)^{-1} \log f_{n_r}(z)$ converges in probability to the constant

$$R \triangleq \lim_{n \to \infty} \frac{I_{n_r}(q)}{n_r} = \lim_{t \to \infty} \frac{I_t(q)}{t} < R(D) + \delta,$$

where we have used inequality (36). It follows from (41) that $\omega^r(\Gamma) \to 0$, which completes the proof of the positive statement.

The negative statement is easily substantiated via proof by contradiction. Indeed, suppose there exists a set $S \subset Y^t$ for which $D(S) \leq D$ but which contains only $N$ elements, where $\log N \leq t[\ln D(D) - \epsilon]$. Let $q^t$ denote the degenerate transition probability that deterministically maps each $x \in X^t$ into whichever $y \in S$ minimizes $\rho_t(x, y)$. Then $D_t(q^t) = D(S) \leq D$, so $q^t \in \xi_t(D)$. Since $S$ contains only $N$ elements, we obtain $R_t(D) \leq \delta_t(q^t) \leq t^{-1} \log N \leq R(D) - \epsilon$. Moreover, this inequality chain must hold for all integral multiples of $t$, too, because repeated use of $q^t$ for successive $t$-blocks leaves the average distortion unchanged and can only decrease the average information rate for a stationary source. The validity of $R_{nt}(D) \leq R(D) - \epsilon$ for arbitrarily large $n$, however, contradicts the definition of $R(D)$.

The practical significance of Theorem 2 resides in the following considerations. If we segment each realization of $[X, \mu]$ into successive $t$-blocks and then map each $t$-block, $x$, so obtained into whichever $y \in S$ minimizes $\rho_t(x, y)$, the distortion that results is $\rho_t(x | S)$ of (18). For stationary $[X, \mu]$ the average value of this distortion will be the same for each successive $t$-block, namely $D(S)$ of (17). This segmenting and
mapping process deterministically associates with each realization \( x \in X \) a particular element of \( Y \), call it \( y_s(x) \). If we let \( v_s \) denote the probability induced on \( Y \) by \([X, \mu]\) and the mapping \( y_s(\cdot) \), then the associated \( t \)-block source \([Y, v_s]\) approximates \([X, \mu]\) with an average distortion of \( D(S) \leq D + \epsilon \). Since each of the approximating \( t \)-blocks that comprise \( y_s(x) \) is an element of the finite set \( S \), \([Y, v_s]\) is a time discrete stationary source with a finite alphabet. It is well known that the entropy per symbol of such a source never exceeds the logarithm of the number of elements in the source alphabet (Fano, 1961). Therefore, Theorem 2 guarantees the existence of a source that approximates \([X, \mu]\) with an average distortion of \( D + \epsilon \) and also possesses an entropy rate that does not exceed \( t^{-1} \log N \leq R(D) + \epsilon \). Now, the channel coding theorem (Shannon, 1948) states that any discrete source with entropy rate \( H \) can, with proper encoding and decoding, be transmitted over any channel of capacity \( C > H \) with an arbitrarily small frequency of errors. (See also Dobrushin, 1959, and Wolfowitz, 1964.) The increase in average distortion that results from transmission of \([Y, v_s]\), over any channel of capacity \( C > R(D) + \epsilon \) therefore can be made arbitrarily small provided \( \rho \) is bounded. (For unbounded \( \rho \) even a single channel error might be disastrous.) Hence, we have

**Corollary 2.** Let \( \epsilon > 0 \) and \( D \geq 0 \) be given. If \([X, \mu]\) is stationary and block ergodic and \( \rho \) is bounded, then it is possible to reproduce the source output at the receiving end of any channel of capacity \( C > R(D) + \epsilon \) with an average distortion with respect to \( F_\rho \), that does not exceed \( D + \epsilon \).

Either, but not both, of the \( \epsilon \)'s in Corollary 2 may be set to zero. The following converse of Corollary 2 also is valid.

**Corollary 3.** If \([X, \mu]\) has rate distortion function \( R(D) \) with respect to \( F_\rho \), then its output cannot be reproduced with an average distortion of \( D \) or less at the receiving end of any memoryless channel of capacity \( C < R(D) \).

**Proof.** Both the proof given by Shannon (1959, Theorem 5) and his subsequent comments regarding the extension to channels with memory can be extended in the usual way to account for abstract alphabets and continuous time. We omit the details.

V. CONCLUSIONS AND EXTENSIONS

In the preceding sections we have formulated and proved the fundamental theorem of rate distortion theory for block ergodic sources with abstract alphabets and single letter fidelity criteria. This is tantamount
to establishing the information stability of the source-fidelity criterion combinations in question, an essential requirement for the principal theorems of the generalized theory of information transmission as given by Dobrushin (1959).

It is possible to extend the above results in several directions. First the extension from single letter distortion measures to distortion measures of span \( g \) (Shannon, 1959) is reasonably straightforward. A distortion measure of span \( g > 0 \) is any nonnegative measurable function \( \eta(z) \) that depends on \( z \) only through \( z^g \). The associated fidelity criterion \( F_n = \{ \eta_t, g \leq t < \infty \} \) is specified, say for continuous time, by

\[
\eta_t(z) = \frac{1}{t-g} \int_0^{t-g} \eta(T^s z) \, ds, \quad t > g.
\]

That is, \( \eta_t(z) \) is the sliding average of \( \eta \) over all successive joint \( g \)-blocks in the joint \( t \)-block, \( z^t \). If \( \eta \) is bounded, this extension is trivial. If not, then it suffices to replace (16) with the assumption that there exists a time index \( t \geq g \) and an element \( b \in Y^t \) such that

\[
\int \eta_{2t}(x, 2b) \, d\mu < \infty,
\]

where \( 2b \in Y^{2t} \) denotes the element obtained by cascading two copies of \( b \).

Another extension concerns the simultaneous imposition of several fidelity criteria, \( F_{\eta_k}, k = 1, 2, \ldots, K \). Let \( \eta_k \) be of span \( g_k \) and suppose that its sliding average is required not to exceed \( D_k \). If we replace \( \eta_k \) by \( n_k' = (D/D_k) \eta_k \), then the average of \( \eta_k' \) must not exceed \( D \) for each \( k \). This condition specifies the set \( \varphi_k(D) \) of permissible transition probabilities \( q^t \), and we continue to define \( R_k(D) \) and \( R(D) \) as in Section III. Condition (16) is replaced by the existence of an element \( b \in Y^t, t \geq g_k \), which simultaneously satisfies (42) for \( k = 1, 2, \ldots, K \).

The block ergodicity assumption clearly may be replaced by the requirement that there exists a divergent sequence \( \{ \tau_i \} \) of positive time indices for which \( [X, \mu] \) is \( \tau_i \)-ergodic. Furthermore, there is good reason to suppose that the results can be extended to arbitrary almost periodic ergodic message sources via still more general versions of Theorem 1.

\[\text{It does not suffice for } \eta_t(x, b) \text{ to be integrable. The objective is to construct a realization } y \in Y \text{ for which } \eta_r(x, y) \text{ is integrable for all } r. \text{ If } y \text{ is to consist of infinite repetitions of some } b \in Y^t, \text{ then it is also necessary not to incur infinite average distortion when the span of length } g \text{ overlaps the end of one } b \text{ and the beginning of the next one.}\]
and Corollary 1 suggested by Perez (1964) on the basis of work by Jacobs (1959); this matter is presently under investigation.

Another point worthy of further study is whether or not condition (16) (or (42)) can be relaxed. For example, if the source and approximating alphabets both are the real line, the source produces independent Cauchy variates, and the distortion measure is mean squared error, then (16) is not satisfied for any choice of $\beta$. $R(D)$ is still defined, however, although there no longer is a distortion value $D_{\text{max}}$ above which it vanishes identically. At present there is no guarantee in such cases that the function $R(D)$ can be meaningfully interpreted as specifying the rate at which the source produces information relative to the fidelity criterion.

**APPENDIX. ERGODICITY, BLOCK ERGODICITY, AND WEAK MIXING**

A block ergodic source is ergodic by definition, but an ergodic source need not be block ergodic. For example, if we concentrate $\mu$ with equal weights of $\frac{1}{2}$ on the two alternating sequences $x = \cdots 0101010 \cdots$ and $Tx$, then $[X, \mu]$ is $\tau$-ergodic if and only if $\tau$ is odd.

A time continuous source $[X, \mathcal{X}, \mu]$ is said to be weakly mixing if for any two sets $E_1, E_2 \in \mathcal{G}$,

$$
\lim_{t \to \infty} \frac{1}{t} \int_0^t | \mu(E_1 \cap T^{-s}E_2) - \mu(E_1)\mu(E_2) | \, ds = 0. \tag{A.1}
$$

(For time discrete sources the integral is replaced by a sum from 0 to $t - 1$.) Following Pinsker (1960, p. 70), we restrict attention to those time continuous sources $[X, \mu]$ for which

$$
\lim_{\tau \to 0} \mu(\Delta T^{-\tau}E) = 0 \tag{A.2}
$$

for every $E \in \mathcal{X}$, where $\Delta = (A \cap \bar{B}) \cup (\bar{A} \cap B)$, and call such sources continuous random processes.

**Lemma.** All weakly mixing continuous random processes are block ergodic.

**Proof.** We must show that any weak mixing source that satisfies (A.2) is $\tau$-ergodic for all $\tau > 0$. Choose $\tau > 0$, let $E \in \mathcal{X}$ be any $\tau$-invariant set, and let $t$ in (A.1) tend to infinity in increments of $\tau$. Then it follows from (A.1) that

$$
0 = \lim_{n \to \infty} \frac{1}{\tau} \sum_{k=0}^{n-1} \int_{k\tau}^{(k+1)\tau} | \mu(E \cap T^{-s}E) - \mu^2(E) | \, ds.
$$

Changing integration variable to $r = s - k\tau$ and noting from $\tau$-in-
variance of $E$ that
\[ E \cap T^{-r}E = E \cap T^{-r}(T^{-kr}E) = E \cap T^{-r}E, \]
we conclude that
\[ 0 = \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} \int_0^r | \mu(E \cap T^{-r}E) - \mu^2(E) | \, dr \]
\[ = \int_0^r | \mu(E \cap T^{-r}E) - \mu^2(E) | \, dr. \] \hspace{1cm} (A.3)

Thus, $\mu(E \cap T^{-r}E) = \mu^2(E)$ for almost all $r \in [0, \tau]$. Since $E \Delta T^{-r}E$ contains $E \cap T^{-r}E$, we have
\[ \mu(E \Delta T^{-r}E) \geq \mu(E) - \mu(E \cap T^{-r}E) = \mu(E) - \mu^2(E) \]
for almost all $r \in [0, \tau]$. It follows from (A.2) that $\mu(E) = \mu^2(E)$ and, hence, that $[X, \mu]$ is $\tau$-ergodic as was to be shown.

If a time discrete source $[X, \mu]$ is weakly mixing, then retaining only every $\tau$th term in the sum corresponding to (A.1) shows that $[X, \mu]$ is also weakly mixing and thus a fortiori ergodic. Hence, weak mixing always implies block ergodicity.

However, block ergodicity does not imply weak mixing. To show this we first exhibit a measure space $[\Omega, \mathcal{F}, P]$ on which a $P$-preserving transformation can be defined that fails to be weakly mixing even though all its powers are ergodic. For example, let $\Omega$ be the unit circle in the complex plane, $\mathcal{F}$ be the $\sigma$-field generated by the open arcs, and $P$ be radian Lebesgue measure. If $c \in \Omega$ and $Vc = c^\omega$ defines the transformation $V$, then it is well known that $V$ is ergodic if and only if $c$ is not a root of unity. But if $c$ is not a root of unity, then neither is $c^n$ for any positive integer $n$, so $V^n$ is ergodic for all $n$. However, $V$ is not weakly mixing. To see this, let $f(\omega) = \omega$ and observe that $U^n f = c^n f$, where $U$ is the isometry induced on $L_2(P)$ by $V$. Employing the notation and weak mixing criterion of Halmos (1956, p. 38), we have $(U^n f, f) = c^n$, but $(f, 1) = (1, f) = 0$, so
\[ \lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n-1} | (U^j f, f) - (f, 1)(1, f) | = 1 \neq 0 \]
and therefore $V$ is not weakly mixing.

Next, we establish an isomorphism between the rotation $V$ on $\Omega$ and the shift $T$ on the space $X$ of binary sequences as follows. Let $\phi: \Omega \to X$ take the point $\omega$ into the binary sequence $x = (\cdots, x_{-1}, x_0, x_1, \cdots)$
specified by the condition \( x_n = 0 \) if and only if \( V'\omega \) belongs to the upper half of the unit circle, \( n = 0, \pm 1, \pm 2, \cdots \). By using the fact that the orbit of each \( \omega \in \Omega \) is dense when \( c \) is not a root of unity, it is easy to show that \( \phi \) is one-to-one. However, \( \phi \) is not onto, the sequence comprised entirely of zeros being an example of an \( x \) that is not the image under \( \phi \) of any \( \omega \in \Omega \). Nevertheless, \((\Omega, \mathcal{F}, P, V)\) is isomorphic to \((X, \mathcal{A}, \mu, T)\) in the sense of Billingsley (1965, p. 53) if \( \mu \) is defined by \( \mu(E) = P(\phi^{-1}E) \) and \( \mathcal{A} \) is the \( \sigma \)-field generated by the image of \( \mathcal{F} \) under \( \phi \). Since (weak) mixing and (block) ergodicity are invariants under such an isomorphism, \([X, \mu]\) is a stationary block ergodic source that is not weakly mixing.

We conclude from the above that block ergodicity is more restrictive than ergodicity but less stringent than weak mixing.

ACKNOWLEDGMENT

I am indebted to Dr. I. G. Stiglitz for pointing out the applicability of Lemma 2 which greatly simplifies the random coding argument. I also wish to thank the referee for his highly constructive comments regarding both content and style.


REFERENCES


