The Annals of Probability
1978, Vol. 6, No. 1, 138–143

THE DISTORTION-RATE FUNCTION FOR NONERGODIC SOURCES

BY P. C. SHIELDS, D. L. NEUHOFF, L. D. DAVISSON[1] AND F. LEDRAPPIER

University of Toledo, University of Michigan, University of Southern California and Université Paris VI

The distortion rate function $D(R)$ is defined as an infimum of distortion with respect to a mutual information constraint. The usual coding theorems assert that, for ergodic souces, $D(R)$ is equal to $\delta(R)$, the least distortion attainable by block codes of rate R. If a source has ergodic components $\{\theta\}$ with weighting measure $dw(\theta)$, it has been shown by Gray and Davisson that $\delta(R)$ is the integral of the components $\delta_\theta(R)$ with respect to $dw(\theta)$. We show that $D(R)$ is the infimum of the integrals of $D_\theta(R_\theta)$ where the integral of R_θ is R. Our method of proof also gives a formula for the $\bar{d}$-distance in terms of ergodic components and shows that $D(R) = D'(R)$, which is defined as the infimum of distortion subject to an entropy constraint.

1. Introduction. For our purposes a source is a stationary process with a finite alphabet $A = \{a_1, a_2, \cdots, a_k\}$. We define

$$x^n = (x_0, \cdots, x_{n-1}); \qquad A^n = \{x^n \mid x_i \in A, 0 \leqq i \leqq n-1\}$$

and the distortion measures

$$d(a_i, a_j) = 0 \quad \text{if} \quad i = j; \qquad d(a_i, a_j) = 1 \quad \text{if} \quad i \neq j;$$

$$d(x^n, y^n) = \frac{1}{n} \sum_{i=0}^{n-1} d(x_i, y_i).$$

A source $x = \{X_N\}$ defines a measure μ_x on A^n by the formula

$$\mu_x(x^n) = \operatorname{Prob}(X^n = x^n), \quad x^n \in A^n.$$

This gives the entropy functions

$$H(X^n) = -E_x(\log \mu_x(X^n))$$
$$H(x) = \lim n^{-1} H(X^n)$$

where E_x denotes conditional expectation with respect to μ_x.

If $(x, y) = \{X_n, Y_n\}$ is a joint process with alphabet $A \times A$ then the conditional entropy is

$$H(Y^n \mid X^n) = -E_{x,y}\left(\log\left(\frac{\mu_{x,y}(X^n, Y^n)}{\mu_x(X^n)}\right)\right)$$

and the mutual information is

$$I(X^n, Y^n) = H(Y^n) - H(Y^n \mid X^n).$$

Received June 14, 1976; revised May 17, 1977.

[1] Work of the first three authors supported by NSF Grants MP 575-08324, ENG 75-10207 and ENG 74-23990.

AMS 1970 *subject classifications*. Primary 94A20; Secondary 28A70.

Key words and phrases. Distortion rate function, entropy, $\bar{d}$-distance.

The distortion rate function $D_x(R)$ is defined as

$$D_x(R) = \inf_n \inf_{\nu \in P_x(R,n)} E_\nu(d(X^n, Y^n))$$

where $P_x(R, n)$ is the class of all measures ν on $A^n \times A^n$ such that $\mu_x(x^n) = \sum_{y^n} \nu(x^n, y^n)$, $x^n \in A^n$ and such that if y is the process defined by $\sum_{x^n} \nu(x^n, y^n)$, $y^n \in A^n$ then $(1/n)I(X^n, Y^n) \leqq R$.

If x and y are sources then the $\bar{d}$-distance between them is defined by

$$\bar{d}(x, y) = \inf_{z \in x \vee y} E_z(d(X_0, Y_0))$$

where $x \vee y$ is the class of all stationary processes z with alphabet $A \times A$ which have x and y as marginals. The function $D_x'(R)$ is defined by

$$D_x'(R) = \inf_{H(y) \leqq R} \bar{d}(x, y) .$$

The ergodic decomposition of a source x is described as follows: there is a probability space, (Φ_x, Σ_x, w_x) such that for each $\theta \in \Phi_x$ there is an ergodic source x_θ such that for each $x^n \in A^n$ the function $\theta \to \mu_\theta(x^n)$ is Σ_x-measurable (where $\mu_\theta = \mu_{x_\theta}$) and

$$\mu_x(x^n) = \int \mu_\theta(x^n)\, dw_x(\theta) .$$

The existence of such a decomposition is well known (see [4, 11]).

Our main theorems are

THEOREM 1. $D_x(R) = D_x'(R)$.

THEOREM 2. $D_x(R) = \inf \int D_\theta(R_\theta)\, dw_x(\theta)$ *where this infimum is taken over all* Σ_x-*measurable functions* $\theta \to R_\theta$ *for which* $\int R_\theta\, dw_x(\theta) \leqq R$.

THEOREM 3. $\bar{d}(x, y) = \inf \int \bar{d}(x_\theta, y_\phi)\, \mathrm{dr}\, (\theta, \phi)$, *where this infimum is taken over all measures r on the product space* $\Phi_x \times \Phi_y$ *which have* w_x *and* w_y *as marginals.*

The proofs of these results will be accomplished by a sequence of lemmas.

LEMMA 1. *If x is ergodic then* $D_x(R) = D_x'(R)$.

This was proved by Gray, Neuhoff and Omura [5].

LEMMA 2. $D_x(R) \leqq D_x'(R)$.

This was also proved in [5].

LEMMA 3. $D_x(R) \geqq \inf \int D_\Theta(R_\Theta)\, dw_x(\Theta)$, *where this infimum is over all measurable functions* $\Theta \to R_\Theta$ *for which* $\int R_\Theta\, dw_x(\Theta) \leqq R$.

To prove this, replace $D_x(R)$ and $D_\Theta(R_\Theta)$ by their nth order approximations $D_x^n(R)$ and $D_\Theta^n(R_\Theta)$ (see [5]). Choose $\nu \in P_x(R, n)$ so that $E_\nu(d(X^n, Y^n)) \leqq D_x^n(R) + \varepsilon$, and define $q(y^n \mid x^n) = \nu(x^n, y^n)/\mu_x(x^n)$. Now define ν_Θ by $\nu_\Theta(x^n, y^n) = \mu_\Theta(x^n) q(y^n \mid x^n)$ and put $R_\Theta^n = (1/n)I_{\nu_\Theta}(X^n, Y^n)$. Then $\nu_\Theta \in P_{x_\Theta}(R_\Theta^n, n)$ and

$$I_\nu(X^n, Y^n) \geqq \int I_{\nu_\Theta}(X^n, Y^n)\, dw_x(\Theta)$$

since I_{ν_Θ} is concave in μ_Θ (see [3, pages 39f.]). This gives

$$\int R_\Theta^n\, dw_x(\Theta) \leqq R .$$

We also have $\int \nu_\Theta(x^n, y^n)\, dw_x(\Theta) = \nu(x^n, y^n)$ for $(x^n, y^n) \in A^n \times A^n$ so that

$$\begin{aligned}\int D_\Theta{}^n(R_\Theta{}^n)\, dw_x(\theta) &\leqq \int E_{\nu_\Theta}(d(X^n, Y^n))\, dw_x(\Theta)\\ &= E_\nu(d(X^n, Y^n))\\ &\leqq D_x{}^n(R) + \varepsilon\,.\end{aligned}$$

Now take the infimum on n,then let $\varepsilon \to 0$ to obtain Lemma 3.

Our next lemma is a technical result about product spaces and makes use of the following notation. (Ω_1, B_1) and (Ω_2, B_2) will denote copies of the unit interval with Borel sets B_i, m will be a regular Borel probability measure on (Ω_1, B_1), g will be a bounded Borel function on Ω_2 and D a $B_1 \times B_2$ measurable set. We let F_m denote the family of Borel measurable maps $f\colon \Omega_1 \to \Omega_2$ such that

$$m\{\omega_1\colon (\omega_1, f(\omega_1)) \notin D\} = 0\,.$$

That is, the graph of f is m-a.e. contained in D. We also let D_{ω_1}, be the ω_1-cross section, that is, $D_{\omega_1} = \{\omega_2\colon (\omega_1, \omega_2) \in D\}$.

LEMMA 4. $\inf_{f\in F_m} \int g(f(\omega_1))\, dm(\omega_1) = \int \inf_{\omega_2 \in D_{\omega_1}} g(\omega_2)\, dm(\omega_1)$.

This result is a simple consequence of Theorem 6.3 of [8].

We now make use of Lemma 4 to establish

LEMMA 5. *If* $\Theta \to R_\Theta \geqq 0$ *is measurable and* $\int R_\Theta\, dw_x(\Theta) \leqq R$ *then* $D_x'(R) \leqq \int D_\Theta'(R_\Theta)\, dw_x(\Theta)$.

To prove this we let $\Phi(A)$ be the family of all stationary processes with alphabet A. This is a complete separable metric space, hence Borel isomorphic to the unit interval. Here we take the metric on $\Phi(A)$ to be $\bar{d}(x, y) = \sum a_n \bar{d}_n(X^n, Y^n)$, where a_n is a suitable convergence factor and $\bar{d}_n = \sum_{a^n \in A^n} |\mu_x(a^n) - \mu_y(a^n)|$. We then put $\Omega_1 = \Phi_x$, $\Omega_2 = \Phi_x \times \Phi(A)$, $m = w_x$, $g(\theta, y) = \bar{d}(x_\theta, y)$, $D = \{(\theta, (\theta, y)) \mid h(y) \leqq R_\theta\}$, $F_1 = F_m = \{y\colon \theta \to (\theta, y_\theta) \mid h(y_\theta) \leqq R_\theta$, a.e.-$w_x\}$. Lemma 4 then gives

$$\int D_x'(R_\theta)\, dw_x(\theta) = \int \inf_{h(y)\leqq R_\theta} \bar{d}(x_\theta, y)\, dw_x(\theta) = \inf_{y\in F_1} \int \bar{d}(x_\theta, y_\theta)\, dw_x(\theta)\,.$$

Fix $y \in F_1$ and put $\Omega_1 = \Phi_x$, $\Omega_2 = \Phi(A \times A)$, $D = \{(\Theta, z)\colon z \in x_\Theta \vee y_\Theta\}$ and $F_y = F_m$, then apply Lemma 4 to obtain

$$\begin{aligned}\inf_{y\in F_1} \int \bar{d}(x_\Theta, y_\Theta)\, dw_x(\Theta) &= \inf_{y\in F_1} \int \inf_{z\in x_\Theta \vee y_\Theta} E_z(d(X_0, Y_0))\, dw_x(\Theta)\\ &= \inf_{y\in F_1} \inf_{z\in F_y} \int E_{z(\Theta)}(d(X_0, Y_0))\, dw_x(\Theta)\\ &\geqq \inf_{h(y)\leqq \int R_\theta dw_x(\Theta)} (\inf_{z\in x\vee y} E_z(d(X_0, Y_0)))\\ &= D_x'(\textstyle\int R_\Theta\, dw_x(\Theta))\,.\end{aligned}$$

This proves Lemma 5.

Theorems 1 and 2 are now consequences of these lemmas. We have

$$\begin{aligned}D_x(R) \overset{(1)}{\leqq} D_x'(R) &\overset{(2)}{\leqq} \inf_{\int R_\Theta dw_x(\Theta)\leqq R} \int D_\Theta'(R_\Theta)\, dw_x(\Theta)\\ &\overset{(3)}{=} \inf_{\int R_\Theta dw_x(\Theta)\leqq R} \int D_\Theta(R_\Theta)\, dw_x(\Theta) \overset{(4)}{\leqq} D_x(R)\,.\end{aligned}$$

Here (1) uses Lemma 2, (2) uses Lemma 5, (3) uses Lemma 1 and (4) uses Lemma 3. This proves both Theorem 1 and Theorem 2.

To prove Theorem 3 we make use of another property of the Rohlin ergodic decomposition, along with Lemma 4. Suppose x and y are stationary processes with respective ergodic decompositions and weight measures, $\{x_\Theta, w_x(\Theta)\}$ and $\{y_\phi, w_y(\phi)\}$. Let $z \in x \vee y$. Then there is a measure $r \in w_x \vee w_y$ and a measurable mapping $(\Theta, \phi) \to Z_{\Theta,\phi} \in x_\Theta \vee y_\phi$, r-a.e., so that for each n and any set $B \in A^n \times A^n$

$$\mu_z(B) = \int \mu_{z_{\Theta,\phi}}(B) \,\mathrm{dr}\,(\Theta, \phi) \,.$$

Here we use $w_x \vee w_y$ to denote the class of measures on $\Phi_x \times \Phi_y$ with w_x and w_y as marginals. For fixed $r \in w_x \vee w_y$ we let F_r be the set of measurable mappings $z\colon (\theta, \phi) \to z_{\theta,\phi}$ for which $z_{\theta,\phi} \in x_\theta \vee y_\phi$, r-a.e. We therefore have

$$\begin{aligned} \bar{d}(x, y) &= \inf_{z \in x \vee y} E_z(d(X_0, Y_0)) \\ &= \inf_{r \in w_x \vee w_y} (\inf_{z \in F_r} \textstyle\int E_{z_{\theta,\phi}}(d(X_0, Y_0)) \,\mathrm{dr}\,(\theta, \phi)) \qquad (5)\\ &= \inf_{r \in w_x \vee w_y} \textstyle\int \inf_{z \in x_\theta \vee y_\phi} E_z(d(X_0, Y_0)) \,\mathrm{dr}\,(\theta, \phi) \\ &= \inf_{r \in w_x \vee w_y} \textstyle\int \bar{d}(x_\theta, y_\phi) \,\mathrm{dr}\,(\theta, \phi) \,. \end{aligned}$$

This proves Theorem 3. The equality (5) is obtained by using Lemma 4 with $\Omega_1 = \Phi_x \times \Phi_y$, $m = r$, $\Omega_2 = \Phi(A \times A)$, $D = \{((\Theta, \phi), z)\colon z \in x_\theta \vee y_\phi\}$ and $g(z) = E_z(d(X_0, Y_0))$.

REMARK 1. If $\delta_z(R)$ is the optimal performance achievable by block codes of rate R, then Gray and Davisson [4] have established the result

$$\delta_x(R) = \int \delta_\theta(R) \, dw(\theta) \,. \qquad (6)$$

Since $D_x(R) \leqq \delta_x(R)$ with equality for ergodic sources x it follows that

$$D_x(R) \leqq \int D_\theta(R) \, dw(\theta) \,.$$

Our results show that in general this inequality is strict. Kieffer [7] has also established (6) by a more direct argument than used in [4].

If $\delta_x'(R)$ is the optimal performance achievable by sliding block codes of rate R, then Gray, Neuhoff and Ornstein [6] have shown that

$$\delta_x'(R) = D_x'(R)$$

holds for aperiodic sources. Furthermore, in [6] it has been shown that for ergodic sources $\delta_x'(R) = \delta_x(R)$. (See also [12].) Our results show, therefore, that for aperiodic sources

$$\delta_x'(R) = D_x(R) \,.$$

REMARK 2. Theorem 3 allows one to give a proof that the class of all Markov chains with alphabet A is separable in the $\bar{d}$-metric. First note that the Friedman–Ornstein proof [2] that mixing Markov chains are finitely determined shows that such chains with rational transition probabilities are $\bar{d}$-dense in the class of mixing chains. If x is an ergodic but nonmixing chain with matrix P and periodic classes $\{C_1, C_2, \cdots, C_d\}$ then we define $\bar{x}$ as the chain with states $C_1 \times C \times \cdots \times C_d$ and transition probabilities.

$$P_{(i_1,i_2,\cdots,i_d),(j_1,j_2,\cdots,j_d)} = P_{i_d j_1} P_{j_1 j_2} \cdots P_{j_{d-1} j_d} .$$

The obvious coding from $\bar{x}$ sequences to x-sequences is not stationary but does map typical strings into typical strings so that if y has the same periodic classes then

$$\bar{d}(x, y) \leqq \bar{d}(\bar{x}, \bar{y}) .$$

Since $\bar{x}$ is mixing this shows that the class of ergodic chains with rational entries is $\bar{d}$-dense in the class of all ergodic chains.

Suppose now that x and y are nonergodic chains with the same ergodic classes $\{G_1, G_2, \cdots, G_l\}$. Let $p_x(G_i)$ and $p_y(G_i)$ denote the respective probabilities that a state belongs to G_i and let x^i and y^i denote the respective restrictions to G_i. Obviously $\bar{d}(x^i, y^j) = 1$ if $i \neq j$ so for any weighting w which has μ_x and μ_y as marginals we have according to Theorem 3,

$$\bar{d}(x, y) \leqq \sum \bar{d}(x^i, y^i) w_{ii} + \sum_{i \neq j} w_{ij} .$$

Thus if the entries in the matrix of x are close enough to the entries in the matrix of y then by our above argument for the ergodic case, we know that each $\bar{d}(x^i, y^i)$ will be small. If furthermore each $p_x(G_i)$ is close to $P_y(G_i)$ we conclude that $\bar{d}(x, y)$ will be small. This completes the proof that the class of all Markov chains with a given finite alphabet A is $\bar{d}$-separable. This $\bar{d}$-separability enables one to establish various universal coding results for the class of all chains [10, 13].

REMARK 3. Our basic results, Theorems 1, 2 and 3, were first announced by the second author for the case of finite ergodic decompositions [9]. The first and second author worked out lengthy proofs of these results. The third and fourth authors provided the much simpler proofs contained in this paper.

REFERENCES

[1] BREIMAN, L. (1960). On achieving channel capacity in finite memory channels. *Illinois J. Math.* **4** 246–252.

[2] FRIEDMAN, N. and ORNSTEIN, D. (1970). On isomorphism of weak Bernoulli transformations. *Advances in Math.* **5** 365–394.

[3] GALLAGER, R. (1968). *Information Theory and Reliable Communications.* Wiley, New York.

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

[5] GRAY, R., NEUHOFF, D. and OMURA, J. (1975). Process definitions of distortion-rate functions and source coding. *IEEE Trans. Information Theory* **21** 524–532.

[6] GRAY, R., NEUHOFF, D. and ORNSTEIN, D. (1975). Nonblock source coding with a fidelity criterion. *Ann. Probability* **3** 478–491.

[7] KIEFFER, J. (1975). On the optimum average distortion attainable by a fixed rate coding of a nonergodic source. *IEEE Trans. Information Theory* **21** 190–193.

[8] MACKEY, G. W. (1967). Borel structures in groups and their duals. *Trans. Amer. Math. Soc.* **85** 134–165.

[9] NEUHOFF, D. (1975). Optimal source coding for stationary nonergodic sources. 13th Annual Allerton Conf. Circuits and System Theory. Urbana, Illinois.

[10] NEUHOFF, D., GRAY, R. and DAVISSON, L. (1975). Fixed rate universal block source coding with a fidelity criterion. *IEEE Trans. Information Theory* **21** 511–523.

[11] ROHLIN, V. A. (1949). On the fundamental ideas of measure theory. *Mat. Sb.* **25** 107–150. AMS Translations, Ser. 1, 10 (1962), 1–54.

[12] SHIELDS, P. and NEUHOFF, D. (1977). Block and sliding-block source coding. *IEEE Trans. Information Theory* **23** 211–215.

13] SHIELDS, P. and NEUHOFF, D. (1977). Fixed-rate universal codes for Markov sources. *IEEE Trans. Information Theory*. To appear.

P. C. SHIELDS
DEPARTMENT OF MATHEMATICS
THE UNIVERSITY OF TOLEDO
TOLEDO, OHIO 43606

L. D. DAVISSON
DEPARTMENT OF ELECTRICAL ENGINEERING
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIFORNIA 90007

D. L. NEUHOFF
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN 48109

F. LEDRAPPIER
LABORATOIRE DE PROBABILITÉ
UNIVERSITÉ PARIS VI
9 QUAI SAINT BERNARD
75320 PARIS CEDEX 05
FRANCE