The Annals of Mathematical Statistics
1971, Vol. 42, No. 6, 2079–2125

EPSILON ENTROPY AND DATA COMPRESSION[1]

EDWARD C. POSNER AND EUGENE R. RODEMICH
Jet Propulsion Laboratory, California Institute of Technology

0. Summary. This article studies efficient data transmission, or "data compression", from the standpoint of the theory of epsilon entropy. The notion of the entropy of a "data source" is defined. This quantity gives a precise measure of the amount of channel capacity necessary to describe a data source to within a given fidelity, epsilon, with probability one, when each separate "experiment" must be transmitted without storage from experiment to experiment. We also define the absolute epsilon entropy of a source, which is the amount of capacity needed when storage of experiments is allowed before transmission. The absolute epsilon entropy is shown to be equal to Shannon's rate distortion function evaluated for zero distortion, when suitable identifications are made. The main result is that the absolute epsilon entropy and the epsilon entropy have ratio close to one if either is large. Thus, very little can be saved by storing the results of independent experiments before transmission.

1. Introduction. An area of information theory that is becoming more and more important is now known as "Data Compression". This is the theory of efficient handling of data for transmission or storage. As the space program goes to more and more distant missions, such as those to the outer planets, the importance of using the spacecraft-to-Earth communication link efficiently becomes more and more important. And here on Earth, as our technology gets more and more computer based, more and more data is being transmitted between different locations, so that the economic importance of data compression is increasing rapidly.

A need has been felt by workers in data compression to have a mathematical theory of what they are trying to do. It is the purpose of this paper to provide such a theory. Thus, we start off by abstracting the notion of the source of the data to be transmitted as a *probabilistic metric space*. Such a space is a triple (X, d, μ) with the following two properties:

(1) (X, d) is a complete separable metric space, of points of the set X under the metric d.

(2) (X, B, μ) is a probability space, where B is the Borel field generated by the open sets of X; i.e., B is the field of *Borel sets* of X. (However, we will speak of subsets of X as measurable if they belong to the completion of B under μ.)

Received February 13, 1970; revised May 14, 1971.

[1] This paper presents the results of one phase of research carried out at the Jet Propulsion Laboratory, California Institute of Technology, under Contract No. NAS 7-100, sponsored by the National Aeronautics and Space Administration.

The points of X represent the possible experimental outcomes, i.e., the data; the metric d is a "fidelity criterion", such that $d(x, y)$ is the loss of fidelity when one outcome x occurs, but another outcome y is thought to have occurred.

We were given an $\varepsilon \geqq 0$, as the allowed loss of fidelity. But we are also given a $\delta \geqq 0$, such that the loss of fidelity can exceed ε, but only with probability δ or less. Any physical data transmission system has a probability $\delta > 0$ associated with it, such that, with probability δ, the system does not work as specified. For example, an "analog-to-digital converter" has as its δ the "off-scale" probability.

Given a probabilistic metric space (X, d, μ), which we abbreviate as X, we wish to transmit outcomes within ε or less with probability $1-\delta$ or more. What can this mean? One reasonable interpretation is the following. If, given a received message, we know the actual outcome within ε, then we know that the actual outcome falls within an ε-set, i.e., within a (measurable) set of diameter at most ε. However, with probability δ or less, there is no transmission and no such set of diameter at most ε. We shall sometimes also use a definition involving spheres of radius $\varepsilon/2$ or less.

Now it might appear that "mixed strategies" should be considered, in which the set of diameter ε depends on other "things", or random variables, not depending on past or future samples from X, in addition to depending on the actual outcome. However, it turns out, and is not hard to show, that there is a "pure strategy" at least as good as any mixed strategy, insofar as reducing the load on the communication channel. A "pure strategy" assigns a given outcome to a fixed ε-set. Thus, in a pure strategy, these ε-sets are to be disjoint. That is, we have a collection of disjoint ε-sets, such that with probability $1-\delta$ or more, a point of the space X lies in one of the ε-sets. This leads to the definition of $\varepsilon;\delta$ *partition.*

DEFINITION. Given $\varepsilon \geqq 0$, $\delta \geqq 0$, an $\varepsilon;\delta$ partition of the probabilistic metric space X is a finite or denumerably infinite partition of part of X by (disjoint) ε-sets, such that the union of the sets in the partition has probability $1-\delta$ or more.

A data compression system based on a given $\varepsilon;\delta$ partition U of X works as follows: observe an outcome, see into which set of the partition U the outcome falls, and merely transmit information which uniquely determines that set. In fact, any ("pure strategic") data transmission system which yields fidelities of ε or better with probability $1-\delta$ or more is of this type.

(Other authors [10, 4] usually consider restricting some average distance to be ε or better, instead of the actual distance between actual outcome and transmitted outcome. The actual distance is felt to be the appropriate criterion in judging most data systems.)

So we now have an $\varepsilon;\delta$ partition U of X; how much of a load does this imply for our communication system? A measure of the load is the entropy of the partition, $H(U)$. This entropy is the number of bits necessary to transmit information as to which set of U the outcome fell, given that the outcome was in some set of U. Thus, using Shannon's formula [11], we define $H(U)$ as follows.

$$\text{Let } U = \{U_i\}, \text{ with } \mu(U_i) = p_i, \sum p_i = p \geqq 1-\delta.$$

Let $q_i = \frac{p_i}{p}$, so that $\{q_i\}$ is a probability distribution.

Then the entropy $H(U)$ is defined as the entropy of $\{q_j\}$:

$$H(U) = \sum q_i \log \frac{1}{q_i}. \tag{1}$$

Thus, $H(U)$ is nonnegative, and can be equal to $+\infty$.

The interpretation of $H(U)$ is the number of bits per sample (we use base e logarithms, even though we speak of bits instead of "nats") necessary to describe into which set of U the outcome falls, when minimum expected-length binary encoding is used for this purpose, but storage between experiments is not allowed.

However, the partition U may not have been especially well chosen with a view to minimizing $H(U)$. Thus, let $U_{\varepsilon;\delta}$ denote the class of $\varepsilon;\delta$ partitions of X (when $\varepsilon > 0$, this class is non-empty). Then we define $H_{\varepsilon;\delta}(X)$, the *epsilon;delta entropy* of X, as

$$H_{\varepsilon;\delta}(X) = \inf_{U \in U_{\varepsilon;\delta}} H(U). \tag{2}$$

(When $\delta = 0$, we write $H_\varepsilon(X)$, the ε-entropy of X. If $\varepsilon = 0$ <nd $U_{0;\delta}$ is empty, $H_{0;\delta}(X)$ is infinite.)

Thus, the $\varepsilon;\delta$ entropy of X, which may equal $+\infty$ even when $\varepsilon > 0$, provided that $\delta = 0$, measures the minimum number of bits necessary to describe at least $1-\delta$ of X within accuracy ε. When $\delta = 0$, we can say that the ε-entropy of X is the number of bits necessary to describe elements of X within ε (without storage, but we will say more about this later), with probability 1. We are mainly concerned in this paper with ε-entropy; $\varepsilon;\delta$ entropy enters chiefly as a device in proving theorems.

We have chosen to use sets of diameter ε in our definition of epsilon entropy to agree with the definition used in the theory of the epsilon entropy of compact metric spaces [13]. However, other workers in information theory use in effect sets of radius $\varepsilon/2$ in the definition. We will treat both cases; when it is necessary to distinguish the two definitions, we will use the terms "diametric entropy" and "radial entropy", respectively. The original paper on the subject of data compression and epsilon entropy [7] used diametric entropy, although Kolmogorov [2] uses radial entropy.

The purpose of the remainder of this paper is to make the statements of this section more precise. Various techniques will be given for obtaining upper and lower bounds for ε-entropy. The ε-entropy defined here will be related to the epsilon entropy of compact metric spaces and to Shannon's rate distortion theory. Finally, a very precise "channel coding theorem" and its converse will be stated and proved for ε-entropy, which makes the relevance of ε-entropy to data compression clear.

2. Connection with ε-entropy for compact metric spaces. A definition for the ε-entropy of a compact metric space finds wide use in various kinds of approximation theory; see [13] for a good expository treatment and many references.

DEFINITION. An ε-partition ($\varepsilon > 0$) of a compact metric space X with metric d is a partition of all of X by disjoint Borel sets of diameters at most ε (caution: other workers use 2ε instead of ε).

DEFINITION. The *ε-entropy* $K_\varepsilon(X)$ of a compact metric space X is the logarithm of the minimum number of sets in any ε-partition of X (since X is compact, $K_\varepsilon(X)$ is finite).

An information-theoretic interpretation of $K_\varepsilon(X)$ can be given as follows. Suppose data is generated as elements of a compact metric space X. We wish to transmit outcomes of X, not allowing storage from one experiment to the next, outputting binary words of fixed length, using as short a length as possible. What is this shortest length? The ε-entropy $K_\varepsilon(X)$ is the answer, except for round-off in the logarithm. If, however, any probabilistic information is known about X, that is, if the data is selected according to a Borel distribution μ on X, the average word length may be able to be shortened by taking advantage of μ in assigning word lengths. That is,

$$H_\varepsilon(X) \leqq K_\varepsilon(X)$$

since the entropy of an ε-partition of X by n sets is bounded from above by $\log n$. In general, one does not consider $K_\varepsilon(X)$ for non-compact metric spaces, since this number can be infinite. (In fact, if $K_\varepsilon(X) < \infty$, all $\varepsilon > 0$, then X is totally bounded.) In the general (non-compact) case, then, words of fixed length cannot be used if one wishes to communicate with probability 1.

Examples can be given (a pentagon with center not connected to one vertex) of a complete separable metric space in which $K_\varepsilon(X)$ is not the supremum of $H_\varepsilon(X)$ taken over all Borel probability measures μ on X, although equality is often obtained.

We remark that the construction of ε-entropy for certain product spaces, done in the next section for probabilistic metric spaces, can also be done for compact metric spaces. Thus, we can speak of the *absolute epsilon entropy* of a compact metric space, which is essentially the infimum of the number of bits necessary to describe all of X to within ε, when words of fixed length are used, but when a large number n of experiments from X are performed before transmission. We shall not go into this matter further in this paper, but will elsewhere [3].

3. Epsilon entropy of product spaces. In this section, we begin the main subject of this paper, which is the relation of ε-entropy to channel coding theorems and their converse. We first define the kind of finite products of probabilistic metric spaces we consider. This definition is motivated by the desire to know each outcome of a sequence of experiments to within ε.

DEFINITION. Let $X = (X, d, \mu)$ and $Y = (Y, e, \nu)$ be probabilistic metric spaces. Then the *product* (*probabilistic metric*) space $X \times Y = (X \times Y, \max(d, e), \mu \times \nu)$ is the probabilistic metric space whose point set is the Cartesian product $X \times Y$ of

X and Y, whose metric f is defined by $f[(x, y), (x', y')] = \max [d(x, x'), e'(y, y')]$, and whose measure is the product measure of μ and ν.

DEFINITION. Let X be a probabilistic metric space. Define $X^{(1)} = X$, $X^{(n+1)} = X^{(n)} \times X$, $n \geqq 1$. Thus $X^{(n)}$ is the n-fold product of X with itself.

Since each $X^{(n)}$ is a probabilistic metric space, we can study the sequence $H_\varepsilon(X^{(n)})$. We have the following lemma.

LEMMA 1. *The sequence* $\{H_\varepsilon(X^{(n)})\}$ *is subadditive in n. That is,*

$$H_\varepsilon(X^{(m+n)}) \leqq H_\varepsilon(X^{(m)}) + H_\varepsilon(X^{(n)}). \tag{3}$$

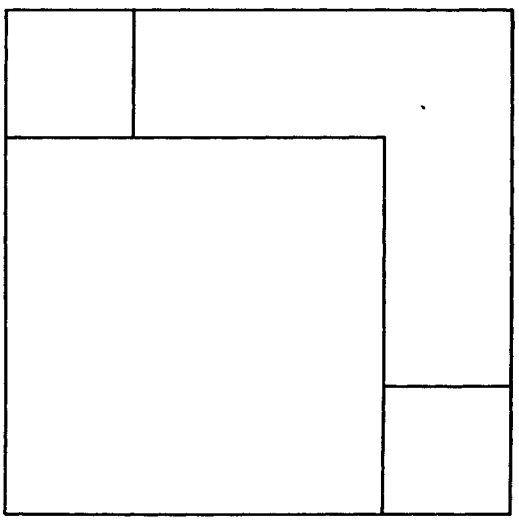

FIG. 1. $\frac{3}{4}$–partition of $X \times X$.

PROOF. By [8a], Lemma 8, if we have an ε-partition U of $X^{(m)}$ and U' of $X^{(n)}$, then the partition $U \times U'$ of $X^{(m+n)}$, defined as the partition consisting of all products of sets in U with sets in U', has the property that

$$H(U \times U') = H(U) + H(U'). \tag{4}$$

By the definition of the metric on $X^{(m+n)}$, $U \times U'$ is an ε-partition of $X^{(m+n)}$. Since we can demand $H(U) = H_\varepsilon(X^{(m)})$, $H(U') = H_\varepsilon(X^{(n)})$ ([7], Theorem 2), we have an ε partition V of $X^{(m+n)}$ with

$$H(V) = H_\varepsilon(X^{(m)}) + H_\varepsilon(X^{(n)}). \tag{5}$$

The lemma follows.

We observe that $H_\varepsilon(X^{(m+n)})$ can be less than $H_\varepsilon(X^{(m)}) + H_\varepsilon(X^{(n)})$. For example, let $m = n = 1$, and let X be the unit interval with Lebesgue measure and linear metric; let $\varepsilon = \frac{3}{4}$. Then $2H_\varepsilon(X) = 2((\frac{3}{4}) \log (\frac{4}{3}) + (\frac{1}{4}) \log 4)$, whereas $H_\varepsilon(X \times X) = (\frac{9}{16}) \log (\frac{16}{9}) + (\frac{5}{16}) \log (\frac{16}{5}) + 2(\frac{1}{16}) \log 16 < 2H_\varepsilon(X)$; in fact, it can be proved that the $\frac{3}{4}$ partition of $X \times X$ in Figure 1 achieves $H_\varepsilon(X \times X)$; diametric entropy or radial entropy can be used here.

Another inequality which will be useful is

$$H_\varepsilon(X^{(m)}) \leqq H_\varepsilon(X^{(m+n)}). \tag{6}$$

This is a special case of the following lemma:

LEMMA 2. *Let X and Y be separable metric spaces with metrics d_1 and d_2, and let $Z = X \times Y$ have the metric d_3, such that if $z_j = (x_j, y_j)$, then*

$$d_3(z_1, z_2) \geqq \max [d_1(x_1, x_2), d_2(y_1, y_2)].$$

Let X, Y, Z be probabilistic metric spaces under some Borel probability distribution on Z and under the induced marginal distributions on X and Y. Thus $H_\varepsilon(X) \leqq H_\varepsilon(Z)$.

PROOF. Let μ_3 be measure on Z and μ_1 measure on X. Take an ε-partition $U = \{U_j\}$ of Z for which

$$H_\varepsilon(Z) = \sum_j \mu_3(U_j) \log \frac{1}{\mu_3(U_j)}.$$

Let the $\{U_j\}$ be arranged in order of decreasing probability. If V_j is the closure of the projection of U_j on X, V_j is a Borel set of diameter $\leqq \varepsilon$, by hypothesis, and $\mu_1(V_j) \geqq \mu_3(U_j)$, since $U_j \subset V_j \times Y$. Similarly,

$$\mu_1(V_1 U V_2 U \cdots U V_j) \geqq \mu_3(U_1 U U_2 U \cdots U U_j).$$

Hence, if we define $W_1 = V_1$, and

$$W_j = V_j - (V_1 U V_2 U \cdots U V_{j-1}), \qquad j \geqq 2,$$

$W = \{W_j\}$ is an ε-partition of X with

$$\textstyle\sum_{k=1}^{j} \mu_1(W_k) \geqq \sum_{k=1}^{j} \mu_3(U_k).$$

It follows [7, Lemma 2] that $H(W) \leqq H(U)$. Therefore $H_\varepsilon(X) \leqq H_\varepsilon(Z)$. The lemma follows.

We define the *absolute ε-entropy* $I_\varepsilon(X)$ as

$$I_\varepsilon(X) = \lim_{n \to \infty} \frac{1}{n} H_\varepsilon(X^{(n)}). \tag{7}$$

After Lemma 1, a well-known result on subadditive sequences [3] implies that this limit exists if the $H_\varepsilon(X^{(n)})$ are all finite, and

$$I_\varepsilon(X) \leqq H_\varepsilon(X), \qquad n = 1, 2, \cdots. \tag{8}$$

Furthermore, Lemma 1 and equation (6) ensure that if any $H_\varepsilon(X^{(n)})$ is finite, they are all finite. If these entropies are all infinite, we interpret (7) to mean $I_\varepsilon(X) = \infty$.

$I_\varepsilon(X)$ is clearly nonnegative, and the following result holds.

LEMMA 3. *$I_\varepsilon(X)$ is zero if and only if $H_\varepsilon(X)$ is zero, and infinite if and only if $H_\varepsilon(X)$ is itself infinite.*

PROOF. After (8) and the remarks above, the only thing to show is that if $H_\varepsilon(X) > 0$, then $I_\varepsilon(X) > 0$.

When $H_\varepsilon(X) > 0$, there is a number $p < 1$ such that any ε-set in X has measure $\leqq p$ [7, Theorem 3]. Any ε-set in $X^{(n)}$ lies in the product of n ε-sets of the component spaces (its n projections), hence has measure at most p^n. It follows that if $U = \{U_j\}$ is any ε-partition of $X^{(n)}$,

$$H(U) = \sum \mu^{(n)}(U_j) \log \frac{1}{\mu^{(n)}(U_j)} \geqq \sum \mu^{(n)}(U_j) \log \frac{1}{p^n} = n \log \frac{1}{p}.$$

Thus $1/n\, H_\varepsilon(X^{(n)}) \geqq \log 1/p$ for all n, which implies $I_\varepsilon(X) \geqq \log 1/p$. Lemma 3 is proved.

The following definition and lemma are needed before we can get down to channel coding theorems.

DEFINITION. $I_{\varepsilon;[\delta]}(X) = \lim_{n\to\infty} \inf n^{-1} H_{\varepsilon;\delta}(X^{(n)})$.

LEMMA 4. *If* $H_\varepsilon(X) < \infty$, $I_\varepsilon(X) = \lim_{\delta\to 0+} I_{\varepsilon;[\delta]}(X)$.

PROOF. The existence of an ε-partition $U = \{U_j\}$ of X with finite entropy will be used to estimate the difference between $H_{\varepsilon;\delta}(X^{(n)})$ and $H_\varepsilon(X^{(n)})$.

Let V be an $\varepsilon;\delta$ partition of $X^{(n)}$ with $H(V) = H_{\varepsilon;\delta}(X^{(n)})$, and let B be the part of $X^{(n)}$ not covered by V. Let $W = \{W_j\}$ be the restriction to B of the product partition $U^{(n)} = \{U_j(n)\}$ of $X^{(n)}$. If B has measure μ_B, the partition WUV of X^n has entropy

$$H(WuV) = \mu_B H(W) + (1-\mu_B)H(V) + \mu_B \log \frac{1}{\mu_B} + (1-\mu_B) \log \frac{1}{1-\mu_B}.$$

We have

$$H(W) = \sum_j \frac{\mu^{(n)}(W_j)}{\mu_B} \log \frac{\mu_B}{\mu^{(n)}(W_j)}.$$

Hence

$$H_\varepsilon(X^{(n)}) \leqq (1-\mu_B)H_{\varepsilon;\delta}(X^{(n)}) + (1-\mu_B) \log \frac{1}{1-\mu_B} + \sum_j \mu^{(n)}(W_j) \log \frac{1}{\mu^{(n)}(W_j)}. \tag{9}$$

We group the W_j into two classes:

(I) $\qquad \mu^{(n)}(W_j) < \delta^{\frac{1}{2}} \mu^{(n)}(U_j^{(n)})$,

(II) $\qquad \mu^{(n)}(W_j) \geqq \delta^{\frac{1}{2}} \mu^{(n)}(U_j^{(n)})$.

If n is sufficiently large, every $U_j^{(n)}$ has measure less than $1/e$. Then

$$\begin{aligned} \sum_{(I)} \mu^{(n)}(W_j) \log \frac{1}{\mu^{(n)}(W_j)} &\leqq \sum_{(I)} \delta^{\frac{1}{2}} \mu^{(n)}(U_j^{(n)}) \log \frac{1}{\delta^{\frac{1}{2}} \mu^{(n)}(U_j^{(n)})} \\ &\leqq \delta^{\frac{1}{2}} \log \frac{1}{\delta^{\frac{1}{2}}} + n\delta^{\frac{1}{2}} H(U), \end{aligned} \tag{10}$$

since $H(U^{(n)}) = nH(U)$. For the second class we have

$$\sum_{(II)} \mu^{(n)}(W_j) \log \frac{1}{\mu^{(n)}(W_j)} \leqq \sum_{(II)} \mu^{(n)}(U_j^{(n)}) \log \frac{1}{\mu^{(n)}(U_j^{(n)})}.$$

Each $U_j^{(n)}$ is the product of a certain n sets of U:

$$U_j^{(n)} = U_{k_1} \times U_{k_2} \times \cdots \times U_{k_n},$$

so we have

$$\sum_{(II)} \mu^{(n)}(W_j) \log \frac{1}{\mu^{(n)}(W_j)} \leqq \sum_{l=1}^{n} \sum_{(II)} \mu(U_{k_1}) \cdots \mu(U_{k_n}) \log \frac{1}{\mu(U_{k_l})}.$$

Since

$$\textstyle\sum_{(II)} \mu^{(n)}(W_j) \leqq \mu_B \leqq \delta^{\frac{1}{2}},$$

we must have

$$\textstyle\sum_{(II)} \mu^{(n)}(U_j^{(n)}) \leqq \delta^{\frac{1}{2}}.$$

Also,

$$\textstyle\sum_{k_l \text{ fixed}} \mu^{(n)}(U_j^{(n)}) = \mu(U_{k_l}).$$

Hence

$$\sum_{(II)} \mu^{(n)}(W_j) \log \frac{1}{\mu^{(n)}(W_j)} \leqq \sum_{l=1}^{n} \sum_{k_l} \min [\delta^{\frac{1}{2}}, \mu(U_{k_l})] \log \frac{1}{\mu(U_{k_l})},$$

or

$$\sum_{(II)} \mu^{(n)}(W_j) \log \frac{1}{\mu^{(n)}(W_j)} \leqq n \sum_{k} \min [\delta^{\frac{1}{2}}, \mu(U_k)] \log \frac{1}{\mu(U_k)}.$$

Combining the last inequality with (9) and (10), and replacing the first terms on the right in (10) by upper bounds,

$$\frac{1}{n} H_\varepsilon(X^{(n)}) \leqq \frac{1}{n} H_{\varepsilon;\delta}(X^{(n)}) + \frac{1}{ne} + \sum_k \min [\delta^{\frac{1}{2}}, \mu(U_k)] \log \frac{1}{\mu(U_k)} + \delta^{\frac{1}{2}} H(U) + \frac{1}{n} \delta^{\frac{1}{2}} \log \frac{1}{\delta^{\frac{1}{2}}}.$$

Take the lower limit as $n \to \infty$:

$$I_\varepsilon(X) \leqq I_{\varepsilon;[\delta]}(X) + \sum_k \min [\delta^{\frac{1}{2}}, \mu(U_k)] \log \frac{1}{\mu(U_k)} + \delta^{\frac{1}{2}} H(U).$$

The series on the right approaches zero as $\delta \to 0$, for it is bounded, term by term, by the series for $H(U)$, and each term approaches zero. Hence

$$\liminf_{\delta \to 0+} I_{\varepsilon;[\delta]}(X) \geqq I_\varepsilon(X).$$

On the other hand, it follows directly from the definitions and the inequality $H_{\varepsilon;\delta}(X^{(n)}) \leqq H_\varepsilon(X^{(n)})$ that $I_{\varepsilon;[\delta]}(X) \leqq I_\varepsilon(X)$. Hence, $I_{\varepsilon;[\delta]}(X) \to I_\varepsilon(X)$ as $\delta \to 0$. Lemma 4 is proved.

Because of needs in the last section, we have the following slight strengthening of Lemma 4.

LEMMA 5. *Let* $H_\varepsilon(X) < \infty$. *Then*

$$I_\varepsilon(X) = \inf_{[\delta_n] \to 0} \liminf_{n \to \infty} \frac{1}{n} H_{\varepsilon;\delta_n}(X^{(n)}). \tag{11}$$

PROOF. Let $\{\delta_n\} \to 0$, and $\delta > 0$ be given. Then for n sufficiently large, $\delta_n \leqq \delta$. Thus,

$$H_{\varepsilon;\delta}(X^{(n)}) \leqq H_{\varepsilon;\delta_n}(X^{(n)}),$$

if n is sufficiently large. Hence, for the given δ,

$$\liminf_{n\to\infty} \frac{1}{n} H_{\varepsilon;\delta}(X^{(n)}) \leqq \liminf_{n\to\infty} \frac{1}{n} H_{\varepsilon;\delta_n}(X^{(n)}),$$

so, by Lemma 4, we conclude

$$I_\varepsilon(X) \leqq \liminf_{n\to\infty} \frac{1}{n} H_{\varepsilon;\delta_n}(X^{(n)}).$$

Consequently

$$I_\varepsilon(X) \leqq \inf_{[\delta_n]\to 0} \liminf_{n\to\infty} \frac{1}{n} H_{\varepsilon;\delta_n}(X^{(n)}). \tag{12}$$

On the other hand,

$$H_{\varepsilon;\delta_n}(X^{(n)}) \leqq H_\varepsilon(X^{(n)}), \tag{13}$$

so that

$$\liminf_{n\to\infty} \frac{1}{n} H_{\varepsilon;\delta_n}(X^{(n)}) \leqq \liminf_{n\to\infty} \frac{1}{n} H_\varepsilon(X^{(n)}). \tag{14}$$

(7) then implies

$$\liminf_{n\to\infty} \frac{1}{n} H_{\varepsilon;\delta_n}(X^{(n)}) \leqq I_\varepsilon(X)$$

for any sequence $\{\delta_n\}$ of nonnegative numbers. In particular,

$$\inf_{[\delta_n]\to 0} \liminf_{n\to\infty} \frac{1}{n} H_{\varepsilon;\delta_n}(X^{(n)}) \leqq I_\varepsilon(X). \tag{15}$$

Equations (15) and (12) taken together complete the proof of Lemma 5.

4. Noisy channel coding theorems. We can now relate our concept of epsilon entropy to channel coding. First note that (11) states that if $H_\varepsilon(X)$ is finite, then $I_\varepsilon(X)$ is the number of bits per sample necessary to describe X to within ε over a noiseless channel with probability approaching 1 when arbitrarily long storage is allowed.

The careful reader will at this juncture observe that storage is necessary, however, to achieve even $H_\varepsilon(X)$. For the "single-use" number of bits necessary is actually not $H_\varepsilon(X)$, but can be arbitrarily close to log 2 less [1, page 71]. However, since we are interested chiefly in the case $H_\varepsilon(X)$ large, we shall ignore this difference, and refer to $H_\varepsilon(X)$ itself as the number of bits necessary when storage between experiments is not allowed.

A noisy channel coding theorem for $I_\varepsilon(X)$ can be stated and proved as in Shannon [12]. Before we can think of stating a precise result, we must review some definitions [5, Chapter 2]. Let X and Y be probability spaces with measures μ and ν, and let ρ be a probability distribution on $X \times Y$ defined on the product σ-field of μ and ν. Further, let

$$\rho(A \times Y) = \mu(A),$$

for A μ-measurable, and let

$$\rho(X \times B) = \nu(B) \quad \text{(or sometimes } \nu_\rho(B)\text{)}$$

for B ν-measurable. Then the *mutual information* of ρ (more precisely, of μ and ν with respect to ρ), $I(\rho)$, is defined as follows. Let

$$\frac{d\rho}{d\mu d\nu}$$

denote the Radon–Nikodym derivative of ρ with respect to product measure $\mu \times \nu$. Then

$$I(\rho) = \infty$$

if ρ is not absolutely continuous with respect to $\mu \times \nu$, that is, if $d\rho/d\mu\, d\nu$ is infinite on a set of positive $\mu \times \nu$-measure. If, on the other hand, ρ is absolutely continuous with respect to $\mu \times \nu$, then

$$(16) \qquad I(\rho) = E_\rho\left[\log\left(\frac{d\rho}{d\mu d\nu}\right)\right] = \int_{X \times Y} \left(\log \frac{d\rho}{d\mu d\nu}\right) d\rho$$

$$= \int_{X \times Y} \left(\log \frac{d\rho}{d\mu d\nu}\right)\left(\frac{d\rho}{d\mu d\nu}\right) d\mu d\nu,$$

which may still be infinite. However, since we have

$$(17) \qquad -\frac{1}{e} \leqq t \log t, \qquad \text{for} \quad t \geqq 0,$$

$I(\rho)$ is well defined as a real number or $+\infty$. And since the function $t \log t$ is convex in $t \geqq 0$, we have, from (16),

$$I(\rho) \geqq \left(\log\left[\int_{X \times Y} \left(\frac{d\rho}{d\mu d\nu}\right) d\mu d\nu\right]\right)\left(\int_{X \times Y} \left(\frac{d\rho}{d\mu d\nu}\right) d\mu d\nu\right).$$

Since

$$\int_{X \times Y} d\rho = 1,$$

we conclude

$$(18) \qquad I(\rho) \geqq 0.$$

It turns out to be easier to first consider the case of radial entropy. The diametric case will be done afterwards with a more complicated definition which will be related to the radial definition. Let $R_\varepsilon(X)$ be defined as that class of probability distributions ρ on $X \times X$ such that:

(i) $\rho(A \times X) = \mu(A)$ for A a Borel set in X;

(ii) $\rho(\{(x, y)/d(x, y) \leqq \varepsilon/2\}) = 1$.

Then the noisy channel coding theorem in question turns out to be the statement that

$$I_\varepsilon(X) = \inf_{\rho \in R_\varepsilon(X)} I(\rho) = I_\varepsilon'(X) \quad \text{say.} \tag{19}$$

This result will now be shown to follow from Shannon's proof in [12], coupled with our definitions and results. We will now explain what is going on, under the added restriction that $H_\varepsilon(X)$ be finite. This restriction will be removed in the next section.

In his Rate Distortion paper [12], Shannon studies coding for a source X with a fidelity criterion F which, for our purposes, is a nonnegative Borel function of the distance d between two points of X, which is a non-decreasing function of distance. He then asks the question of how much channel capacity is needed to send independent outcomes of X over a given noisy channel, if it is desired to keep the average distortion D between the actual outcome x and the decoded outcome y bounded by the constant A. The distortion D is defined as the expectation of $F(x, y)$, the expected loss of fidelity between actual outcome and decoded value. In our case,

$$\begin{aligned} F(x, y) &= 0, \qquad d(x, y) \leqq \varepsilon/2; \\ F(x, y) &= 1, \qquad d(x, y) > \varepsilon/2. \end{aligned} \tag{20}$$

It is desired to know if there is a value C_0 of channel capacity such that, over any channel with capacity $C > C_0$, outcomes of X can be transmitted over the channel in such a way that the distortion is zero, with probability approaching 1, whereas, if $C < C_0$, outcomes cannot be transmitted over the channel in such a way that the distortion is zero, with probability approaching 1, as the block length becomes infinite. Shannon's Theorems 1 and 2 of [12] show that this result is indeed true, where

$$C_0 = \inf_{\rho \in R_\varepsilon(X)} I(\rho). \tag{21}$$

We state the desired result in a lemma. We mention that although the proof is adapted from Shannon's Theorem 1, the result does not follow from Shannon's statement.

LEMMA 6. *Let* $H_\varepsilon(X) < \infty$. *Then*

$$I_\varepsilon'(X) = I_\varepsilon(X).$$

In any case,

$$I_\varepsilon'(X) \leqq I_\varepsilon(X).$$

PROOF. First we shall prove that $I_\varepsilon(X) \leqq I_\varepsilon'(X)$. Suppose that $I_\varepsilon'(X) < \infty$. Let δ, η be positive numbers. We shall show that (39) is valid for η sufficiently large (so that (24), (31), (35) are (37) are satisfied) and use this to show $I_\varepsilon(X) \leqq I_\varepsilon'(X)$.

By the definition of $I_\varepsilon'(X)$, we can choose $\rho_\varepsilon R_\varepsilon(X)$ such that

$$I(\rho) \leqq I_g'(X)+\eta. \tag{22}$$

Let $v = v_\rho$ be the second marginal distribution of ρ on $X \times X$. For a positive n, we consider the space Z_n which is the product of n independent copies of $X \times X$,

$$Z_n = (X \times X)^{(n)}$$

with measure $\rho^{(n)}$, the product of the measure ρ on each $X \times X$. Z_n can also be written as

$$Z_n = X^{(n)} \times X^{(n)} \tag{23}$$

where the first factor has product measure $\mu^{(n)} = \mu \times \mu \times \cdots \times \mu$, and the second factor has product measure $v^{(n)}$. Also, $\mu^{(n)}$ and $v^{(n)}$ are the marginal distributions of $\rho^{(n)}$ in this factorization.

Let $\bar{x} = (x_1, \cdots, x_n)$, $\bar{y} = (y_1, \cdots, y_n)$ be the coordinates of a point of Z_n in the representation (23). Consider the function

$$J(\bar{x}, \bar{y}) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{d\rho(x_i, y_i)}{d\mu(x_i)dv(y_i)}.$$

By the weak law of large numbers, for the given δ there is an n_0 such that

$$n \geqq n_0 \tag{24}$$

implies that on a set Γ in Z_n of probability $\geqq 1-\delta^2$ we have

$$J(\bar{x}, \bar{y}) \leqq I(\rho)+\eta. \tag{25}$$

Because of condition (ii) in the definition of $R_\varepsilon(X)$, we can demand that

$$d^{(n)}(\bar{x}, \bar{y}) \leqq \varepsilon/2, \qquad (\bar{x}, \bar{y}) \in \Gamma.$$

From (25), we have

$$\frac{d\rho^{(n)}(\bar{x}, \bar{y})}{d\mu^{(n)}(\bar{x})dv^{(n)}(\bar{y})} \leqq \exp\left[n(I(\rho)+\eta)\right], \qquad (x, y) \in \Gamma. \tag{26}$$

For any $\bar{x} \in X^{(n)}$, define

$$I_{\bar{x}} = \{\bar{y} \mid (\bar{x}, \bar{y}) \in \Gamma\}.$$

The quantity

$$\rho_{\bar{x}}^{(n)}(L_{\bar{x}}) = \int_{L\bar{x}} \frac{d\rho^{(n)}(\bar{x}, \bar{y})}{d\mu^{(n)}(\bar{x})dv^{(n)}(\bar{y})}\, dv^{(n)}(\bar{y}) \tag{27}$$

is the probability of $L_{\bar{x}}$ given $\bar{x}$, and

$$\int_{X^{(n)}} \rho_{\bar{x}}^{(n)}(I_{\bar{x}})d\mu^{(n)}(\bar{x}) = \rho^{(n)}(\Gamma) \geqq 1-\delta^2. \tag{28}$$

Let A be the set of $\bar{x}$ such that

$$\rho_{\bar{x}}^{(n)}(L_{\bar{x}}) \geqq 1-\delta, \qquad \bar{x} \in A. \tag{29}$$

It follows easily from (28) that

$$\mu^{(n)}(A) \geqq 1-\delta. \tag{30}$$

Using (26) to estimate the integral in (27) we get

$$\rho_{\bar{x}}^{(n)}(L_{\bar{x}}) \leqq \nu^{(n)}(L_{\bar{x}}) \exp\left[n(I(\rho)+\eta)\right].$$

Combining with (29)

$$\nu^{(n)}(L_{\bar{x}}) \geqq (1-\delta) \exp\left[-n(I(\rho)+\eta)\right], \qquad \bar{x} \in A.$$

Suppose that n is so large that

$$e^{n\eta}(1-\delta) \geqq 1. \tag{31}$$

Then

$$\nu^{(n)}(L_{\bar{x}}) \geqq \exp\left[-n(I(\rho)+2\eta)\right], \qquad \bar{x} \in A. \tag{32}$$

Let N be a positive integer. Choose N independent points $\bar{y}_1, \cdots, \bar{y}_N$ in $X^{(n)}$ according to the distribution $\nu^{(n)}$. Let P_N be the expected $\mu^{(n)}$-probability of the part of $X^{(n)}$ not covered by the union of the spheres $S_{\varepsilon/2}(\bar{y}_j), j = 1, \cdots, N$. We have

$$\begin{aligned} P_N &= E_{\bar{y}_1,\cdots,\bar{y}_N}\left\{\int_{X^{(n)}}\left[1-\chi_{\bar{x}}\left(\bigcup_{i=1}^{N} S_{\varepsilon/2}(\bar{y}_i)\right)\right]d\mu^{(n)}(\bar{x})\right\} \\ &= \int_{X^{(n)}} E_{\bar{y}_1,\cdots,\bar{y}_N}\left\{\prod_{i=1}^{N}\left[1-\chi_{\bar{x}}(S_{\varepsilon/2}(\bar{y}_i))\right]\right\}d\mu^{(n)}(\bar{x}). \end{aligned}$$

By the independence, the integrand is equal to

$$\prod_{i=1}^{N} E_{\bar{y}_i}\left[1-\chi_{\bar{y}_i}(S_{\varepsilon/2}(\bar{x}))\right] = \prod_{i=1}^{N}\left[1-\nu^{(n)}(S_{\varepsilon/2}(\bar{x}))\right].$$

Thus,

$$P_N = \int_{X^{(n)}}\left[1-\nu^{(n)}(S_{\varepsilon/2}(\bar{x}))\right]^N d\mu^{(n)}(\bar{x}).$$

Since $L_{\bar{x}} \in S_{\varepsilon/2}(\bar{x})$, (30) and (32) imply

$$P_N \leqq \delta+(1-\delta)\{1-\exp\left[-n(I(\rho)+2\eta)\right]\}^N.$$

Applying the inequality $1-t \leqq e^{-t} (0 \leqq t \leqq 1)$, we get

$$P_N \leqq \delta+\exp\left\{-N \exp\left[-n(I(\rho)+2\eta)\right]\right\}. \tag{33}$$

Let n be the integer for which

$$N-1 < \exp\left[n(I(\rho)+3\eta)\right] \leqq N. \tag{34}$$

Then from (33) we have $P_N \leqq \delta+\exp(-e^{n\eta})$.

For n sufficiently large,

$$\exp(-e^{n\eta}) < \delta. \tag{35}$$

Then

$$P_N < 2\delta. \tag{36}$$

Suppose that

$$n\eta > \log 2. \tag{37}$$

Then by virtue of the inequality

$$\log(1+e^t) = t+\log(1+e^{-t}) \leqq t+\log 2, \qquad t \geqq 0,$$

(34) implies

$$\log N \leqq n[I(\rho)+4\eta]. \tag{38}$$

From the definition of P_N, it follows that $\bar{y}_1, \cdots, \bar{y}_N$ can be selected so that the set

$$B = \bigcup_{i=1}^{N} S_{\varepsilon/2}(\bar{y}_i)$$

has $\mu^{(n)}$-measure at least $1-P_N$. The set B can be partitioned into N sets of radius $\varepsilon/2$ (constructed from the $S_{\varepsilon/2}(\bar{y}_i)$). Hence, by (36) and (38)

$$H_{\varepsilon;2\delta}(X^{(n)}) \leqq n[I_\varepsilon(X)+4\eta],$$

where $\mu^{(n)}$-measure is understood, or, from (22),

$$H_{\varepsilon;2\delta}(X^{(n)}) \leqq n[I_\varepsilon'(X)+5\eta]. \tag{39}$$

From the definition of $I_{\varepsilon;[\delta]}(X)$ and Lemma 4, (39) implies that if $H_\varepsilon(X) < \infty$,

$$I_\varepsilon(X) \leqq I_\varepsilon'(X)+5\eta.$$

Now let $\eta \to 0$. This shows that

$$I_\varepsilon(X) \leqq I_\varepsilon'(X),$$

as promised.

The reverse inequality is easier: given a partition U of $X^{(n)}$ by sets U_i, of radius $\varepsilon/2$ and centers $\bar{y}_i$, of entropy $H(U)$, we proceed as follows. The partition U induces a joint distribution σ on $(X \times X)^{(n)}$ by associating n-tuples $\bar{x}$ with the centers $\bar{y}$ of the sets U_i containing $\bar{x}$. In the obvious notation, we have measures $\rho_1, \rho_2, \cdots, \rho_n$ on $X \times X$ where ρ_i corresponds to the ith coordinate, and each ρ_i is in $R_\varepsilon(X)$. We have, by a simple calculation,

$$I(\bar{x}, \bar{y}) = H(U).$$

We can demand, given $\eta > 0$, that

$$\frac{1}{n} H(U) \leqq I_\varepsilon(X)+\eta,$$

so that

$$\frac{1}{n} I(\bar{x}, \bar{y}) \leqq I_\varepsilon(X)+\eta. \tag{40}$$

On the other hand, simple properties of I yield

$$I(\bar{x}, \bar{y}) \geqq \sum_{i=1}^{n} I(x_i, \bar{y}) \text{ (since the } \{x_i\} \text{ are independent [5, Section 2.2])}$$
$$\geqq \sum_{i=1}^{n} I(x_i, y_i) = \sum_{i=1}^{n} I(\rho_i) \text{ (since } I(A, (B, C)) \geqq I(A, B)).$$

We conclude that for some i, $1 \leqq i \leqq n$, we have

$$I(\rho_i) \leqq \frac{1}{n} I(\bar{x}, \bar{y}),$$

and recall $\rho_i \in R_\varepsilon(X)$. (37) then yields

$$I(\rho_i) \leqq I_\varepsilon(X)+\eta.$$

Definition (23) yields

$$I_\varepsilon'(X) \leqq I_\varepsilon(X)+\eta,$$

for all $\eta > 0$. We finally conclude

$$I_\varepsilon'(X) \leqq I_\varepsilon(X) \tag{41}$$

which completes the proof of Lemma 6.

The restriction that $H_\varepsilon(X) < \infty$ will be removed when we prove that $H_\varepsilon(X)$ is finite if $I_\varepsilon'(X)$ is finite.

To prove the channel coding result from Lemma 6, first we show that if K is a memoryless channel of capacity C, with

$$C > I_\varepsilon(X),$$

then we can communicate over K keeping the distance between every received outcome and its corresponding transmitted outcome at most $\varepsilon/2$ with probability approaching 1 as the block length increases. To do this, merely take a partition U of $X^{(n)}$ with

$$C > H(U) \geqq I_\varepsilon(X).$$

Use the centers of the U_i as a new source of entropy $H(U)$ and transmit over K with probability of (word) error approaching zero.

To prove the converse, suppose, as in [12], Theorem 1, that we have a block code which encodes all of the n-tuples $\bar{x}$ into m-tuples $\bar{w}$, and suppose that $\bar{w}$ is received as say the m-tuple $\bar{z}$. Suppose also that we have a decoding procedure which decodes m-tuples $\bar{z}$ into n-tuples $\bar{y}$, such that $d^{(n)}(\bar{x}, \bar{y}) \leqq \varepsilon/2$ with probability at least $1-\eta$. Shannon's result uses the condition

$$\frac{1}{n} \sum_{i=1}^{n} d(x_i, y_i) \leqq \frac{\varepsilon}{2}(1+\eta), \qquad \text{with } \eta > 0,$$

but we want the stronger condition with $\eta = 0$. Simple properties of I yield

$$I(\bar{w}, \bar{z}) \geqq I(\bar{x}, \bar{y}).$$

Also, if $\bar{w}_k$, $\bar{z}_k$ denote the first k components of $\bar{w}$, $\bar{z}$ respectively, we have

$$\begin{aligned}\Delta_k(I(\bar{w}, \bar{z})) &= I(\bar{w}, \bar{z}_{k+1}) - I(\bar{w}, \bar{z}_k) \\ &= I(\bar{w}, (\bar{z}_k, z_{k+1})) - I(\bar{w}, \bar{z}_k) \\ &= I(z_{k+1}, (\bar{w}, \bar{z}_k)) - I(z_{k+1}, \bar{z}_k).\end{aligned}$$

Since the channel K is memoryless,

$$I(z_{k+1}, (\bar{w}, \bar{z}_k)) = I(z_{k+1}, w_{k+1}).$$

Thus

$$\begin{aligned}\Delta_k I(\bar{w}, \bar{z}) &= I(z_{k+1}, w_{k+1}) - I(z_{k+1}, \bar{z}_k) \\ &\leqq I(z_{k+1}, w_{k+1}) = I(w_{k+1}, z_{k+1}).\end{aligned}$$

But the channel capacity C per channel letter is defined as sup $I(w, v)$ over all encodings w, decodings v. Hence

$$\Delta_k I(\bar{w}, \bar{z}) \leqq C', \qquad 1 \leqq k \leqq m,$$

where C' is the capacity of K per coded letter. Therefore

$$I(\bar{w}, \bar{z}) \leqq mC', \quad \frac{1}{n} I(\bar{x}, \bar{y}) \leqq \frac{m}{n} C' = C,$$

where C is the capacity per source letter.

Finally, by Lemma 5, since η is arbitrary > 0, $C \geqq I_\varepsilon(X)$, as promised. We shall not go into further detail.

To do the diametric case involves a slight complication in the definition of $R_\varepsilon(X)$. This happens because we can no longer choose a point y to send when we observe x, but instead must send an ε-set Y containing x when x occurs.

So we define matters as follows. Consider the complete separable metric space $Y(X)$ of closed subsets of X under the Hausdorff metric [9, Section 1.4], or rather its closed subspace $Y_\varepsilon(X)$ of sets of diameters at most ε. Then the diametric $R_\varepsilon(X)$ is the class of Borel probability distributions ρ on $X \times Y_\varepsilon(X)$ such that

(i) $\rho(W \times Y_\varepsilon(X)) = \mu(W)$ for Borel $W \subset X$;

(ii) $\rho(\{x, Y\} \mid x \in Y) = 1$.

In other words, ρ sends points of X into closed ε-sets containing the point, with ν_ρ-probability 1.

We then define

$$I_\varepsilon'(X) = \inf_{\rho \in R_\varepsilon(X)} I(\rho),$$

and the proof goes through as before, with

$$\nu_\rho(B) = \rho(X \times B) \quad \text{for Borel} \quad B \subset Y_\varepsilon(X).$$

The question can be raised as to what happens in the radial case if we define $R_\varepsilon(X)$ as above, using the closed subset $Z_\varepsilon(X)$ of $Y_\varepsilon(X)$ of closed sets of radius $\varepsilon/2$.

The answer is that $I_\varepsilon'(X)$ is unchanged. The main point is that unique centers of the $\varepsilon/2$ spheres called out in the definition can be chosen in a measurable fashion. We omit details.

5. Inequalities for I_ε and H_ε. This section gives some inequalities on I_ε and H_ε, and, in particular, proves that $H_\varepsilon(X) < \infty$ if $I_\varepsilon(X) < \infty$, closing a gap in the last section. We remark that all previous lower bounds for $H_\varepsilon(X)$ were lower bounds for $I_\varepsilon(X)$, and, conversely, all previous upper bounds for $I_\varepsilon(X)$ were actually upper bounds for $H_\varepsilon(X)$ ([7], [8b]). The reason it is so hard to get bounds for one and not the other is that, as we shall see, $H_\varepsilon(X)$ and $I_\varepsilon(X)$ are close in ratio if either is large. We shall have to distinguish between radial and diametric entropy some of the time; when necessary, we use ${}_DI_\varepsilon(X)$, ${}_DH_\varepsilon(X)$ for the diametric case, and ${}_RI_\varepsilon(X)$, ${}_RH_\varepsilon(X)$ for the radial case. Observe the trivial inequalities

$$
\begin{aligned}
{}_RH_{2\varepsilon}(X) &\leqq {}_DH_\varepsilon(X) \leqq {}_RH_\varepsilon(X), \\
{}_RI_{2\varepsilon}(X) &\leqq I_\varepsilon(X) \leqq {}_RI_\varepsilon(X).
\end{aligned} \tag{42}
$$

When a result is true for either definition, we shall omit the presubscript. Our next lemma is quite useful, but will be strengthened later. It provides an upper bound on $I_\varepsilon(X)$, but not on $H_\varepsilon(X)$.

LEMMA 7.

$$
I_\varepsilon'(X) \leqq E_\mu\left[\log \frac{1}{\mu(S_{\varepsilon/2}(x))}\right].
$$

PROOF. It is enough to prove the result for ${}_RI_\varepsilon'(X)$, in view of (42). We shall find a $\rho \in R_\varepsilon(X)$ such that

$$
I(\rho) \leqq E_\mu\left[\log \frac{1}{\mu(S_{\varepsilon/2}(x))}\right] \tag{43}
$$

to prove the lemma. This ρ is, roughly speaking, the measure which takes an outcome x in X, and spreads its probability throughout $S_{\varepsilon/2}(x)$ according to μ restricted to $S_{\varepsilon/2}(x)$. That is, for A a Borel subset of $X \times X$,

$$
\rho(A) = \int \frac{\mu(S_{\varepsilon/2}(x) \cap A_x)}{\mu(S_{\varepsilon/2}(x))} d\mu(x), \tag{44}
$$

where

$$
A_x = \{y \mid (x, y) \in A\}.
$$

The definition makes sense, since

$$
\mu(\{x \mid \mu(S_{\varepsilon/2}(x)) = 0\}) = 0.
$$

Note that for B Borel in X,

$$
\nu_\rho(B) = \rho(X \times B) = \int_X \frac{\mu(S_{\varepsilon/2}(x) \cap B)}{\mu(S_{\varepsilon/2}(x))} d\mu(x). \tag{45}
$$

We now evaluate

$$I(\rho) = \int_{X \times X} \log\left(\frac{d\rho}{d\mu dv_\rho}\right) d\rho.$$

Observe that

$$\frac{dv_\rho(y)}{d\mu(y)} = \int_{S\varepsilon/2(y)} \frac{d\mu(x)}{\mu(S_{\varepsilon/2}(x))} = q(y), \qquad \text{say.} \tag{46}$$

A simple calculation shows

$$\frac{d\rho(x, y)}{d\mu(x)dv_\rho(y)} = \frac{1}{q(y)} \frac{\chi_x(S_{\varepsilon/2}(y))}{\mu(S_{\varepsilon/2}(x))}. \tag{47}$$

Now the support of ρ is contained in

$$\{(x, y) \mid \chi_x(S_{\varepsilon/2}(y)) = 1\}.$$

Hence, continuing from (47), we have

$$\begin{aligned} I(\rho) &= \int_{X \times X} \log \frac{1}{\mu(S_{\varepsilon/2}(x))}\, d\rho(x, y) + \int_{X \times X} \log \frac{1}{q(y)}\, d\rho(x, y) \\ &= \int_X \log \frac{1}{\mu(S_{\varepsilon/2}(x))}\, d\mu(x) + \int_X \log \frac{1}{q(y)}\, dv_\rho(y). \end{aligned}$$

So, from (46), we have

$$I(\rho) + E_\mu\left[\log \frac{1}{\mu(S_{\varepsilon/2}(x))}\right] + \int_X q(y) \log \frac{1}{q(y)}\, d\mu(y). \tag{48}$$

Now from (46) again

$$\begin{aligned} \int q(y)d\mu(y) &= \iint \frac{\chi_x(S_{\varepsilon/2}(y))}{\mu(S_{\varepsilon/2}(x))}\, d\mu(x)d\mu(y) \\ &= \int \frac{1}{\mu(S_{\varepsilon/2}(x))} \left[\int \chi_y(S_{\varepsilon/2}(x))d\mu(y)\right] d\mu(x), \end{aligned}$$

so that

$$\int q(y)d\mu(y) = 1. \tag{49}$$

Convexity of $t \log 1/t$ then shows

$$\int q(y) \log \frac{1}{q(y)}\, d\mu(y) \leqq 0. \tag{50}$$

This, coupled with (48), proves the lemma.

The prime on $I_\varepsilon'(X)$ will be removed after the main theorem. The next lemma is sometimes a useful lower bound on $I_\varepsilon(X)$, and, of course, on $H_\varepsilon(X)$.

LEMMA 8. *Let*

$$\alpha = \sup_{S \text{ an } \varepsilon\text{-set}} \mu(S),$$

$$\beta = \sup_{X \in X} \mu(S_{\varepsilon/2}(x)),$$

so that $\beta \leqq \alpha$ *(by* [7], *Theorem* 3, *these* sup's *are* max's*). Then*

$$(51) \qquad {}_DI_\varepsilon(X) \geqq \log \frac{1}{\alpha},$$

$$(52) \qquad {}_RI_\varepsilon(X) \geqq \log \frac{1}{\beta},$$

PROOF. We shall do the ${}_DI_\varepsilon(X)$ case, the other being similar. If ${}_DI_\varepsilon(X) = \infty$, there is nothing to prove. If ${}_DH_\varepsilon(X) < \infty$, let n be a positive integer, and let U be an ε-partition of $X^{(n)}$. Every set U_i in U has probability

$$\mu^{(n)}(U_i) \leqq \alpha^n,$$

by hypothesis. Hence

$$H(U) \geqq \log \frac{1}{\alpha^n},$$

$$H_\varepsilon(X^{(n)}) \geqq \log \frac{1}{\alpha^n},$$

$$\frac{1}{n} H_\varepsilon(X^{(n)}) \geqq \log \frac{1}{\alpha}.$$

Reference to (7) completes the proof of the lemma.

The next lemma is useful to bound the entropy of a union of spaces; it was implicitly used in Lemma 4. By abuse of language, we shall speak of the epsilon entropy of subsets of X even if they are merely measurable and not closed. The interpretation is quite natural: to send X say whether you are in Y or Z, then send Y or Z.

LEMMA 9. *Let* $X = \bigcup_k Y_k$, Y_k *disjoint and measurable. Regard each* Y_k *as a probabilistic metric space in its own right, with measure and metric inherited from* X. *The renormalised probability measure on* Y_k *is* $\mu_k = p_k{}^{-1}\mu$, *where* $\mu(Y_k) = p_k$. *Define*

$$\mathscr{P} = \{p_k\}, \qquad H(\mathscr{P}) = \sum p_k \log \frac{1}{p_k},$$

the entropy of $\mathscr{P}$. *Then*

$$(53) \qquad H_\varepsilon(X) \leqq \sum p_k H_\varepsilon(Y_k) + H(\mathscr{P}),$$

$$I_\varepsilon'(X) \leqq \sum p_k I_\varepsilon'(Y_k) + H(\mathscr{P}).$$

Furthermore, these inequalities become equalities if, for each k, l with $k \neq l$,

$$\inf_{y_k \in Y_k, y_l \in Y_l} d(y_k, y_l) > \varepsilon. \tag{54}$$

PROOF. If $V_k = \{V_{ik}\}$ is an ε-partition of Y_k with $H(V_k) = H_\varepsilon(Y_k)$, let U be the ε-partition of X given by

$$U = \{V_{ik}, \text{ all } i, k\}. \tag{55}$$

Then

$$\begin{aligned} H_\varepsilon(X) &\leqq \sum_{i,k} \mu(V_{ik}) \log \frac{1}{\mu(V_{ik})} \\ &= \sum_k p_k \sum_i \frac{\mu(V_{ik})}{p_k} \log \frac{p_k}{\mu(V_{ik})} + \sum_k p_k \log \frac{1}{p_k} \\ &= \sum_k p_k H_\varepsilon(Y_k) + H(\mathscr{P}), \end{aligned}$$

as required. When (54) is satisfied, every ε-partition of X is of the form (55), and so the assertion about equality holds.

To prove the result for $I_\varepsilon'(X)$, let $\rho_k \in R_\varepsilon(Y_k)$ with

$$I(\rho_k) \leqq I_\varepsilon'(Y_k) + \eta/2^k,$$

η arbitrary > 0. Let ρ be the measure on $X \times X$ such that

$$\rho(A) = \sum_k p_k \rho_k(A \cap (Y_k \times Y_k)). \tag{56}$$

Then $\rho \in R_\varepsilon(X)$. Also,

$$\begin{aligned} v_\rho(B) = \rho(X \times B) &= \sum p_k \rho_k(Y_k \times (B \cap Y_k)) \\ &= \sum p_k v_{\rho_k}(B \cap Y_k). \end{aligned}$$

Thus,

$$\begin{aligned} \frac{d\rho}{d\mu dv_\rho} &= \frac{1}{p_k} \frac{1}{d\mu_k dv_{\rho_k}} \quad \text{on} \quad Y_k \times Y_k, \qquad k = 1, 2, \cdots, \\ &= 0 \quad \text{outside} \quad \bigcup_k (Y_k \times Y_k). \end{aligned} \tag{57}$$

So

$$\begin{aligned} I(\rho) = E_\rho\left(\log \frac{d\rho}{d\mu dv_\rho}\right) &= \sum_k p_k E_{\rho_k}\left[\log\left(\frac{1}{p_k}\left(\frac{d\rho_k}{d\mu_k dv_{\rho_k}}\right)\right)\right] \\ &\leqq \sum_k p_k I_\varepsilon'(Y_k) + H(\mathscr{P}) + \eta. \end{aligned}$$

Since η is arbitrary, the result follows. When (54) holds, every ρ in $R_\varepsilon(X)$ is of the form (56). This proves the lemma.

REMARK. The primes will be removed from $I_\varepsilon'(X)$ in (53) after we prove the main theorem.

6. Packings of sets. This section prepares for the main theorem by introducing a new kind of random coding argument and using it to prove a result on "asymptotic close packing". The argument was first used in Lemma 6.

LEMMA 10. *Let V be the random ε-partition of X defined in the following way. Choose a sequence $\{x_i\}$ of points of X randomly $d\mu$ and independently. Define*

$$W_i = S_{\varepsilon/2}(x_i), \qquad i \geqq 1.$$

Define the ε-partition $V = \{V_n\}$ by

$$V_n = W_n - \bigcup_{i=1}^{n-1} W_i, \qquad n \geqq 1.$$

Then

$$\mu\left(\bigcup_{n=1}^{\infty} V_n\right) = 1 \quad \textit{with probability} \quad 1. \tag{58}$$

Also,

$$H_\varepsilon(X) \leqq E(H(V)) \leqq E\left(\log \frac{1}{\mu(S_{\varepsilon/2}(x))}\right) + E\left(\log \frac{1}{1-\mu(S_{\varepsilon/2}(x))}\right) \cdot E\left(\frac{1-\mu(S_{\varepsilon/2}(x))}{\mu(S_{\varepsilon/2}(x))}\right), \tag{59}$$

and

$$E(\mu(V_n)) = \int \mu(S_{\varepsilon/2}(x))(1-\mu(S_{\varepsilon/2}(x)))^{n-1} d\mu(x). \tag{60}$$

PROOF. We have

$$E(H(V)) = \sum_{n \geqq 1} E\left(\mu(V_n) \log \frac{1}{\mu(V_n)}\right); \tag{61}$$

by convexity,

$$E(H(V)) \leqq \sum_{n \geqq 1} E(\mu(V_n)) \log \frac{1}{E(\mu(V_n))}. \tag{62}$$

But

$$\sum_{i=1}^{n} E(\mu(V_i)) = E\left(\sum_{i=1}^{n} \mu(V_i)\right) = E\mu\left(\bigcup_{i=1}^{n} W_i\right)$$
$$E\left(\int \left[1 - \prod_{i=1}^{n} (1-\chi_x(W_i))\right] d\mu(x)\right)$$

which, by the independence of the x_i, can be written as

$$\int \left[1 - \prod_{i=1}^{n} (1 - E_{x_i}(\chi_x(S_{\varepsilon/2}(x_i)))\right] d\mu(x)$$
$$= \int \left[1 - \prod_{i=1}^{n} (1 - E_{x_i}(\chi_{x_i}(S_{\varepsilon/2}(x))))\right] d\mu(x)$$
$$= \int \left[1 - \prod_{i=1}^{n} (1 - \mu(S_{\varepsilon/2}(x)))\right] d\mu(x).$$

That is,

$$E\left(\mu\left(\bigcup_{i=1}^{n} W_i\right)\right) = \sum_{i=1}^{n} E(\mu(V_i)) = 1 - \int (1-\mu(S_{\varepsilon/2}(x)))^n d\mu(x). \tag{63}$$

Observe that the random variable $1-\mu(S_{\varepsilon/2}(x))$ takes the value 1 with probability zero; therefore

$$\int (1-\mu(S_{\varepsilon/2}(x)))^n d\mu(x) \to 0 \qquad \text{as} \quad n \to \infty, \tag{64}$$

so

$$E(\mu(\textstyle\bigcup_{n=1}^{\infty} V_n)) = 1,$$

and hence

$$\mu(\textstyle\bigcup_{n=1}^{\infty} V_n) = 1$$

with probability 1, which proves (58). So we do indeed have a random family of ε-partitions V of X, whose entropies have expectation $E(H(V))$. Then there is at least one V with entropy at most $E(H(V))$, which proves the first inequality of (59).

Finally, (63) shows that (60) holds. The entropy $H(\{E(\mu(V_n))\})$ of the sequence of probabilities $\{E(\mu(V_n))\}$ is therefore given by

$$\begin{aligned} &H(\{E(\mu(V_n))\}) \\ &\quad = \textstyle\sum_{n\geqq 1} \left[\int \mu(S_{\varepsilon/2}(x))(1-\mu(S_{\varepsilon/2}(x)))^{n-1} d\mu(x)\right] \log \\ &\qquad \times \frac{1}{\int \mu(S_{\varepsilon/2}(x))(1-\mu(S_{\varepsilon/2}(x)))^{n-1}}\, d\mu(x). \end{aligned}$$

By convexity of the function $\log(1/t)$, we conclude

$$\begin{aligned} &H(\{E(\mu(V_n))\}) \\ &\quad \leqq \sum_{n\geqq 1} \left[\int \mu(S_{\varepsilon/2}(x))(1-\mu(S_{\varepsilon/2}(x)))^{n-1} d\mu(x)\right] \\ &\qquad \cdot \int \log \frac{1}{\mu(S_{\varepsilon/2}(x))(1-\mu(S_{\varepsilon/2}(x)))^{n-1}}\, d\mu(x). \end{aligned} \tag{65}$$

(65) yields

$$\begin{aligned} H(\{E(\mu(V_n))\}) = &\left[\int \log \frac{1}{\mu(S_{\varepsilon/2}(x))}\, d\mu\right] \cdot \int \sum_{n\geqq 1} \mu(S_{\varepsilon/2}(x))(1-\mu(S_{\varepsilon/2}(x)))^{n-1} d\mu \\ &+\left[\int \log \frac{1}{1-\mu(S_{\varepsilon/2}(x))}\, d\mu\right] \\ &\cdot \int \sum_{n\geqq 1} (n-1)\mu(S_{\varepsilon/2}(x))(1-\mu(S_{\varepsilon/2}(x)))^{n-1} d\mu. \end{aligned} \tag{66}$$

Evaluating the sums in (66) and using (62) completes the proof of (59). Lemma 10 is proved.

As a consequence of Lemma 10, we have the following result. Note that Lemma 7 yields

$$I_\varepsilon(X) \leqq \log \frac{1}{\alpha} < \infty$$

if $\mu(S_{\varepsilon/2}(x)) \geqq \alpha > 0$ with probability 1.

LEMMA 11. *Let*

$$\mu(S_{\varepsilon/2}(x)) \geqq \alpha > 0 \quad \textit{with probability} \quad 1.$$

Then $H_\varepsilon(X) < \infty$. *In fact,*

$$H_\varepsilon(X) \leqq \log\frac{1}{\alpha} + \frac{1-\alpha}{\alpha}\log\frac{1}{1-\alpha} < \log\frac{1}{\alpha} + 1. \tag{67}$$

PROOF. (63) yields

$$\sum_{i=1}^{n} E(\mu(V_i)) \geqq 1-(1-\alpha)^n, \tag{68}$$

so that [7, Lemma 2] the entropy of the $\{E(\mu(V_n))\}$ sequence is bounded by the entropy of the sequence $\alpha(1-\alpha)^{n-1}$. Consequently,

$$H_\varepsilon(X) \leqq \sum_{n\geqq 1} \alpha(1-\alpha)^{n-1}\log\frac{1}{\alpha(1-\alpha)^{n-1}}, \tag{69}$$

which yields (67) and proves the lemma.

We are now ready for the asymptotic close packing result.

THEOREM 1. *Let X be a probabilistic metric space such that*

$$\mu(S_{\varepsilon/2}(x)) = \alpha \tag{70}$$

with probability 1. *Then*

$${}_R I_\varepsilon(X) = \log\frac{1}{\alpha} \tag{71}$$

and

$$0 \leqq {}_R H_\varepsilon(X) - {}_R I_\varepsilon(X) < 1. \tag{72}$$

If in addition

$$\mu(A) \leqq \alpha \tag{73}$$

for every measurable set A of diameter at most ε, *then*

$${}_D I_\varepsilon(X) = \log\frac{1}{\alpha} \tag{74}$$

and

$$0 \leqq {}_D H_\varepsilon(X) - {}_D I_\varepsilon(X) < 1. \tag{75}$$

REMARK. The probability α must be positive since $\mu(X) = 1$, so that $H_\varepsilon(X) < \infty$ by Lemma 11. The assumption that $\mu(S_{\varepsilon/2}(x)) = \alpha$ with probability 1 and that no measurable set of diameter at most ε has probability exceeding α is satisfied by spheres, toruses, etc., under the natural measure given by surface hyper-area normalized to 1.

PROOF OF THEOREM. First assume only (70). By Lemma 8,

$$_RI_\varepsilon(X) \geqq \log \frac{1}{\alpha}.$$

By Lemma 7,

$$_RI_\varepsilon(X) \leqq \log \frac{1}{\alpha}.$$

Hence (71) holds. Lemma 11 then implies that (72) holds.

Now let both (70) and (73) hold. Lemma 8 gives

$$_DI_\varepsilon(X) \geqq \log \frac{1}{\alpha}.$$

(42) coupled with (71) gives

$$_DI_\varepsilon(X) \leqq \log \frac{1}{\alpha}.$$

Hence (74) holds. (42), coupled with (71), (74), and (72), yields (75). Theorem 1 is proved.

REMARK. We have seen that $H_\varepsilon(X)-I_\varepsilon(X) < 1$ if $\mu(S_{\varepsilon/2}(x))$ is constant with probability 1. It is not known whether this difference can be arbitrarily large, but we shall prove in the next section

$$H_\varepsilon(X)-I_\varepsilon(X) = o(H_\varepsilon(X)),$$

for $H_\varepsilon(X)$ large.

Discussion of Theorem 1. Theorem 1 is called an "asymptotic close packing" result for the class of spaces to which it applies, for it says that the space can be partitioned by ε-sets in such a way that the resulting partition has entropy less than one more than the entropy of a partition consisting of $1/\alpha$ sets each of probability α. When $\alpha \to 0$, then, the space can be partitioned by a partition practically as "nice" as a close-packing one. Such nice partitions use up "most" of the probability with sets of probability "almost" equal to α. One can use Theorem 1 to state and prove apparently otherwise difficult theorems about partitions of n-spheres, regular graphs, etc. ([6]).

7. The main theorem. In order to get the strongest possible form of the main theorem, we first need the following lemma, a strengthening of the known result ([7], page 54) that if $\{a_n\}$ is a nonnegative sequence summing to 1, such that

$$\sum a_n \log n < \infty,$$

then

$$\sum a_n \log \frac{1}{a_n} < \infty$$

also (the converse is true, as well, but we do not need it here).

LEMMA 12. *For $\alpha, B > 0$, consider all nonnegative sequences $a_1, a_2, \cdots$ with*

$$\sum_{n=1}^{\infty} a_n \leqq 1, \tag{76}$$

$$\sum_{n=1}^{\infty} a_n(\alpha+\log n) \leqq B. \tag{77}$$

Let

$$M(B) = \sup \sum_{n=1}^{\infty} a_n \log \frac{1}{a_n}, \tag{78}$$

over all such sequences.

Let λ be the unique solution on $(1, \infty)$ of the equation

$$e^{-\lambda\alpha}[-\zeta'(\lambda)+\alpha\zeta(\lambda)] = Be, \tag{79}$$

where $\zeta(s)$ is the Riemann Zeta-function. For $B > \alpha$, let σ be the unique solution on $(1, \infty)$ of

$$-\zeta'(\sigma)/\zeta(\sigma) = B-\alpha. \tag{80}$$

Let B_0 be the value of B at which $e^{-\lambda\alpha}\zeta(\lambda) = e$. Then

$$M(B) = \lambda B+e^{-1-\lambda\alpha}\zeta(\lambda), \qquad B \leqq B_0,$$

$$= \sigma(B-\alpha)+\log \zeta(\sigma), \qquad B > B_0.$$

As $B \to \infty$,

$$M(B) = B+\log B+O(1);$$

as $B \to 0$,

$$M(B) = \frac{B}{\alpha}\log\frac{\alpha}{B}+o(B).$$

PROOF. For N a positive integer, define

$$M_N(B) = \sup \sum_{n=1}^{N} a_n \log \frac{1}{a_n}, \tag{78'}$$

under the conditions (76) and (77). Clearly

$$M_N(B) \leqq M_{N+1}(B) \leqq M(B), \qquad N \geqq 1.$$

For any $\varepsilon > 0$, if $\{a_n\}$ is a sequence for which the sum in (78) is greater than $M(B)-\varepsilon$, then for N sufficiently large

$$M_N(B) \geqq \sum_{n=1}^{N} a_n \log \frac{1}{a_n} > M(B)-2\varepsilon.$$

Hence

$$M(B) = \lim_{N\to\infty} M_N(B).$$

The region of admissible values of $(a_1, \cdots, a_N)$ in N-space is a compact region. Hence there is a maximizing sequence for the sum in (78'):

$$M_N(B) = \max \sum_{n=1}^{N} a_n \log \frac{1}{a_n}.$$

Under the single condition

$$\sum_{n=1}^{N} a_n \leqq 1, \tag{76'}$$

this sum has a unique local maximum, at $a_n = N^{-1}$, $n = 1, \cdots, N$. Here

$$\sum_{n=1}^{N} a_n(\alpha+\log n) > B,$$

if N is sufficiently large. Considering only such large values of N, we see that the maximizing sequence for $M_N(B)$ must have

$$\sum_{n=1}^{N} a_n(\alpha+\log n) = B. \tag{77'}$$

Suppose first that $\sum_{n=1}^{N} a_n < 1$ for the maximizing sequence. Then $M_N(B)$ is a local maximum of the series in (78') when $a_1, \cdots, a_N$ vary over the region in the hyperplane (77') where $a_n \geqq 0$, $n = 1, \cdots, N$. This maximum clearly does not occur on the boundary of the region, since the function $x \log (1/x)$ has an infinite derivative at $0+$. Hence, by the method of Lagrange multipliers, at the maximum we have

$$\log \frac{1}{a_n} - 1 = \lambda_n(\alpha+\log n), \qquad n = 1, 2, \cdots, N,$$

or

$$a_n = e^{-1-\alpha\lambda_N} n^{-\lambda N}. \tag{81}$$

Define

$$\zeta_N(s) = \sum_{n=1}^{N} n^{-s}.$$

Then if $\{a_n\}$ is given by (81), the condition (76') is

$$e^{-\alpha\lambda_N}\zeta_N(\lambda_N) \leqq e, \tag{82}$$

and (77') becomes

$$e^{-\alpha\lambda_N}[\alpha\zeta_N(\lambda_N) - \zeta_N'(\lambda_N)] = Be. \tag{83}$$

In this case, the value of $M_N(B)$ is

$$\begin{aligned} M_{N1}(B) &= \sum_{n=1}^{N} e^{-1-\alpha\lambda_N} n^{-\lambda_N} [1+\alpha\lambda_N+\lambda_N \log n] \\ &= \lambda_N B + e^{-1-\alpha\lambda_N} \zeta_N(\lambda_N). \end{aligned} \tag{84}$$

Now suppose that equality holds in (76'). Then $M_N(B)$ is the local maximum of the series in (78'), when $a_1, \cdots, a_N$ vary over a region in an $(N-2)$-dimensional hyperplane. Again this maximum cannot occur at a boundary point. Hence, at the maximum,

$$\log \frac{1}{a_n} - 1 = \mu_N + \sigma_N(\alpha+\log n), \qquad n = 1, 2, \cdots, N,$$

or

$$a_n = \exp(-1-\mu_N-\alpha\sigma_N) n^{-\sigma_N}.$$

From (76′) and (77′)

$$\exp(-1-\mu_N-\alpha\sigma_N)\zeta_N(\sigma_N) = 1.$$

Eliminating μ_N, we get

$$\exp(-1-\mu_N-\alpha\sigma_N)[\alpha\zeta_N(\sigma_N)-\zeta_N'(\sigma_N)] = B.$$

$$-\zeta_N'(\sigma_N)/\zeta_N(\sigma_N) = B-\alpha, \tag{85}$$

and

$$a_n = n^{-\sigma_N}/\zeta_N(\sigma_N). \tag{86}$$

Now the value of $M_N(B)$ is

$$\begin{aligned} M_{N2}(B) &= \zeta_N(\sigma_N)^{-1}\sum_{n=1}^{N} n^{-\sigma_N}[\sigma_N \log n + \log \zeta_N(\sigma_N)] \\ &= \sigma_N(B-\alpha)+\log \zeta_N(\sigma_N). \end{aligned} \tag{87}$$

The functions of λ_N, σ_N in (83) and (85) are strictly decreasing on $(-\infty, \infty)$ for $N \geqq 2$, taking all positive values. Thus there is a unique solution λ_N of (83) for any $B > 0$, and a unique σ_N satisfying (85) for any $B > \alpha$.

If (82) is violated, we must have $M_N(B) = M_{N2}(B)$. We shall show that $M_N(B) = M_{N1}$ when (82) is satisfied. Let B_N be the value of B for which

$$e^{-\alpha\lambda_N}\zeta_N(\lambda_N) = e.$$

Then for $B \leqq B_N$, (82) is satisfied, and (83) gives us

$$e^{-\alpha\lambda_N}[\alpha\zeta_N(\lambda_N)-\zeta_N'(\lambda_N)] \geqq Be^{-\alpha\lambda_N}\zeta_N(\lambda_N),$$

or

$$-\zeta_N'(\lambda_N)/\zeta_N(\lambda_N) \geqq B-\alpha.$$

If $B > \alpha$, then σ_N is defined by (85), and we see that

$$\sigma_N \geqq \lambda_N, \tag{88}$$

with equality for $B = B_N$.

Expressing $M_{N1}(B)$ and $M_{N2}(B)$ by (84) and (87), and differentiating, one has

$$\begin{aligned} \frac{d}{dB}[M_{N1}(B)-M_{N2}(B)] &= \lambda_N-\sigma_N+[-B+\alpha-\zeta_N'(\sigma_N)/\zeta_N(\sigma_N)]\frac{d\sigma_N}{dB} \\ &\quad +[B+e^{-1-\alpha\lambda_N}\{\zeta_N'(\lambda_N)-\alpha\zeta_N(\lambda_N)\}]\frac{d\lambda_N}{dB}. \end{aligned}$$

Applying (83) and (85), this reduces to

$$\frac{d}{dB}[M_{N1}(B)-M_{N2}(B)] = \lambda_N-\sigma_N,$$

and by (88), this is nonpositive for $\alpha < B \leqq B_N$. We have $M_{N1}(B_N) = M_{N2}(B_N)$, since at $B = B_N$, $\sigma_N = \lambda_N$ and

$$e^{-\alpha\lambda_N}\zeta_N(\lambda_N) = e^{-\alpha\sigma_N}\zeta_N(\sigma_N) = e.$$

Hence $M_{N1}(B) \geqq M_{N2}(B)$ for $\alpha < B \leqq B_N$.
Now we know that

$$M_N(B) = M_{N1}(B), \qquad B \leqq B_N,$$
$$= M_{N2}(B), \qquad B > B_N.$$

As $N \to \infty$, the functions $\alpha\zeta_N(s) - \zeta_N'(s)$ and $\zeta_N'(s)/\zeta_N(s)$ tend monotonically to the corresponding non-subscripted functions on $(1, \infty)$, while for $s \leqq 1$ they approach $+\infty$ uniformly. Since the limit functions are strictly decreasing on $(1, \infty)$, we must have $\lambda_N \to \lambda$ and $\sigma_N \to \sigma$, the solutions of (79) and (80). Also $B_N \to B_0$.

If $B < B_0$, or $B > B_0$, then for N sufficiently large the same inequality is valid with B_0 replaced by B_N. Hence

$$M(B) = \lim_{n\to\infty} M_{N1}(B) = \lambda B + e^{-1-\lambda\alpha}\zeta(\lambda), \qquad B < B_0,$$
$$= \lim_{n\to\infty} M_{N2}(B) = \sigma(B-\alpha) + \log\zeta(\sigma), \qquad B > B_0.$$

The value of $M(B_0)$ is forced by the obvious monotonicity of the function $M(B)$.

For the asymptotic form of $M(B)$ as $B \to \infty$, note that $\sigma \to 1+$ as $B \to \infty$; hence

$$\zeta(\sigma) = \frac{1}{\sigma-1} + O(1), \qquad -\zeta'(\sigma) = \frac{1}{(\sigma-1)^2} + O(1).$$

From (80)

$$\frac{1}{\sigma-1} = B + O(1), \qquad \sigma = 1 + B^{-1} + O(B^{-2}).$$

Then we have $\zeta(\sigma) = B + O(1)$, and

$$M(B) = (B-\alpha)[1 + B^{-1} + O(B^{-2})] + \log B + O(B^{-1})$$
$$= B + \log B + O(1).$$

As $B \to 0$, $\lambda \to \infty$. Then

$$\zeta(\lambda) = 1 + O(2^{-\lambda}), \qquad -\zeta'(\lambda) = 2^{-\lambda}\log 2 + O(3^{-\lambda}),$$

and by (79),

$$e^{-\lambda\alpha}[\alpha + O(2^{-\lambda})] = Be,$$
$$e^{-\lambda\alpha} = \frac{Be}{\alpha}[1 + O(2^{-\lambda})],$$
$$\lambda = \frac{1}{\alpha}\log\frac{\alpha}{Be} + O(2^{-\lambda}).$$

Hence

$$M(B) = B\left[\frac{1}{\alpha}\log\frac{\alpha}{Be}+O(2^{-\lambda})\right]+\frac{B}{\alpha}[1+O(2^{-\lambda})]$$
$$= \frac{B}{\alpha}\log\frac{\alpha}{B}+o(B).$$

This completes the proof of Lemma 12.

The next lemma gives a lower bound to $I(\rho)$ for every ρ in $R_\varepsilon(X)$.

LEMMA 13. *Let* $\rho \in {}_RR_\varepsilon(X)$. *Then*

$$I(\rho) \geqq E_\mu\left(\log\frac{1}{v_\rho(S_{\varepsilon/2}(x))}\right) = {}_RK(\rho), \quad say. \tag{89}$$

Let $\rho \in {}_DR_\varepsilon(X)$. *Then*

$$I(\rho) \geqq E_\mu\left(\log\frac{1}{v_\rho(A_x)}\right) = {}_DK(\rho) \quad say, \tag{90}$$

where

$$A_x = \{Y \in Y_\varepsilon(X) \mid x \in Y\}.$$

PROOF. Let $\rho \in {}_RR_\varepsilon(X)$. We have

$$I(\rho) = \int\left[\int \log\left(\frac{d\rho}{d\mu dv_\rho}\right)\frac{d\rho}{d\mu dv_\rho}\,dv_\rho\right]d\mu. \tag{91}$$

Note that

$$\int\left(\log\frac{d\rho(x,y)}{d\mu(x)dv_\rho(y)}\right)\frac{d\rho(x,y)}{d\mu(x)dv_\rho(y)}\,dv_\rho(y) = \int_{S_{\varepsilon/2}(x)}\left(\log\frac{d\rho(x,y)}{d\mu(x)dv_\rho(y)}\right)\cdot\frac{d\rho(x,y)}{d\mu(x)dv_\rho(y)}\,dv_\rho(y)$$

by property (ii) of ${}_RR_\varepsilon(X)$.

Hence, by convexity of $t\log t$

$$\int\left(\log\frac{d\rho(x,y)}{d\mu(x)dv_\rho(y)}\right)\frac{d\rho(x,y)}{d\mu(x)dv_\rho(y)}\,dv_\rho(y) \tag{92}$$
$$\geqq v_\rho(S_{\varepsilon/2}(x))\left[\log\int\frac{d\rho(x,y)}{d\mu(x)dv_\rho(y)}\frac{dv_\rho(y)}{v_\rho(S_{\varepsilon/2}(x))}\right]\int_{S_{\varepsilon/2}(x)}\frac{d\rho(x,y)}{d\mu(x)dv_\rho(y)}\frac{dv_\rho(y)}{v_\rho(S_{\varepsilon/2}(x))}.$$

But if

$$\int_{S_{\varepsilon/2}(x)}\frac{d\rho(x,y)}{d\mu(x)dv_\rho(y)}\,dv_\rho(y) = \int_X\frac{d\rho(x,y)}{d\mu(x)dv_\rho(y)}\,dv_\rho(y) = f(x),$$

then

$$\int_B f(x)d\mu(x) = \int\int_{B\times X} d\rho(x,y) = \mu(B),$$

so that $f(x) = 1$ with μ-probability 1. Thus, (92) yields

$$\int\left(\log \frac{d\rho(x, y)}{d\mu(x)dv_\rho(y)}\right)\frac{d\rho(x, y)}{d\mu(x)dv_\rho(y)}\,dv_\rho \geqq \log\frac{1}{v_\rho(S_{\varepsilon/2}(x))}, \tag{93}$$

with μ-probability 1. (91) then proves (89). The proof of (90) is similar and omitted.

The next lemma upper bounds $H_\varepsilon(X)$ in terms of the right-hand side of (89) or (90).

LEMMA 14. *Let $\rho \in {}_R R_\varepsilon(X)$. Then there exists an absolute constant C such that*

$$ {}_R H_\varepsilon(X) \leqq {}_R K(\rho) + \log^+ {}_R K(\rho) + C. \tag{94}$$

Let $\rho \in {}_D R_\varepsilon(X)$. Then, for the same absolute constant C,

$$ {}_D H_\varepsilon(X) \leqq {}_D K(\rho) + \log^+ {}_D K(\rho) + C. \tag{95}$$

PROOF. As in Lemma 10, let V be the random ε-partition of X defined by choosing a sequence $\{X_i\}$ of points of X randomly dv_ρ, defining

$$W_i = S_{\varepsilon/2}(x_i), \qquad i \geqq 1,$$

and then

$$V_n = W_n - \bigcup_{i=1}^{n-1} W_i, \qquad n \geqq 1.$$

The expected $(d\rho^{(n)})$ μ-probability not covered by $\bigcup_{i=1}^n V_i$ is given by

$$E\left[\int \prod_{i=1}^n (1-\chi_x(S_{\varepsilon/2}(x_i)))d\mu(x)\right] = \int (1-v_\rho(S_{\varepsilon/2}(x)))^n\, d\mu(x), \tag{96}$$

as in Lemma 10. Let the set E be defined by

$$E = \{x \mid v_\rho(S_{\varepsilon/2}(x_i)) = 0\}.$$

We claim

$$\mu(E) = 0.$$

For

$$\mu(E) = \rho(E \times X) = \int_E \int_{S_{\varepsilon/2}(x)} \frac{d\mu}{d\mu dv_\rho}\, dv_\rho d\mu = 0,$$

since if $d\rho$ is not absolutely continuous $d\mu\, dv_\rho$, there is nothing to prove.

We conclude that with dv_ρ-probability 1, V is an ε-partition of X under μ-measure, as in Lemma 10. As in (62),

$$E(H(V)) \leqq \sum_{n \geqq 1} E(\mu(V_n)) \log \frac{1}{E(\mu(V_n))}. \tag{97}$$

And, from (96),

$$\sum_{i=1}^n E(\mu(V_i)) = 1 - \int (1-v_\rho(S_{\varepsilon/2}(x)))^n d\mu. \tag{98}$$

Define

$$G(s) = \mu\{x \mid 1 - v_\rho(S_{\varepsilon/2}(x)) < s\},$$

and define

$$m_n = \int (1-v_\rho(S_{\varepsilon/2}(x)))^n d\mu = \int_0^1 s^n dG(s), \tag{99}$$

the nth moment of s distributed $dG(s)$. As in Lemma 10, (97) yields, since $(1-m_1)\log 1/(1-m_1) \leqq 1/e$,

$$(100)\qquad {}_RH_\varepsilon(X) \leqq \sum_{n=1}^{\infty}(m_n-m_{n+1})\log\frac{1}{m_n-m_{n+1}}+(1-m_1)\log\frac{1}{1-m_1}$$

$$\leqq \sum_{n=1}^{\infty}(m_n-m_{n+1})\log\frac{1}{m_n-m_{n+1}}+\frac{1}{e}.$$

On the other hand

$${}_RK(\rho) = E_\mu\left[\log\frac{1}{\nu_\rho(S_{\varepsilon/2}(x))}\right] = \int_0^1 \log\frac{1}{1-s}\,dG(s).$$

Expanding

$$\log\frac{1}{1-s} = s+\frac{s^2}{2}+\frac{s^3}{3}+\cdots,$$

and then integrating, we conclude

$$(101)\qquad {}_RK(\rho) = \sum_{n=1}^{\infty}\frac{m_n}{n}.$$

Let us sum the right-hand side of (101) by parts to obtain, since $m_n \to 0$,

$$\sum_{n=1}^{\infty}\frac{m_n}{n} = \sum_{n=1}^{\infty}\frac{1}{n}\sum_{i=n+1}^{\infty}(m_{i-1}-m_i) = \sum_{i=2}^{\infty}(m_{i-1}-m_i)\sum_{n=1}^{i-1}\frac{1}{n}$$

$$\geqq \sum_{i=1}^{\infty}(m_i-m_{i+1})(\log i+\gamma),$$

where γ is Euler's constant. Hence

$$(102)\qquad {}_RK(\rho) \geqq \textstyle\sum_{i=1}^{\infty}(m_i-m_{i+1})(\log i+\gamma).$$

Now

$$\textstyle\sum_{i=1}^{\infty}(m_i-m_{i+1}) = m_1 \leqq 1:$$

Lemma 12 then gives

$$(103)\qquad \sum_{i=1}^{\infty}(m_i-m_{i+1})\log\frac{1}{m_i-m_{i+1}} \leqq {}_RK(\rho)+\log^+{}_RK(\rho)+C',$$

for some universal constant C'. (100) next yields

$$(104)\qquad {}_RH_\varepsilon(X) \leqq {}_RK(\rho)+\log^+{}_RK(\rho)+C'+1/e.$$

Define $C'+1/e = C$ to prove (94) from (104). The proof of (95) is similar and omitted. This proves Lemma 14.

The next lemma is not really needed for our main purposes, but is inserted to show that $H_\varepsilon(X)$ must approach 0 if $I_\varepsilon(X)$ does.

LEMMA 15. *There exists a universal constant D such that, if $K(\rho) \leqq \frac{1}{4}(1-1/e)^2$,*

$$H_\varepsilon(X) \leqq D(K(\rho))^{\frac{1}{2}}.$$

PROOF. We shall improve the inequality (100) by getting an upper bound for

$$(1-m_1)\log\frac{1}{1-m_1}$$

in terms of $K(\rho)$. From definition (89), we find

$$\mu\{x \mid \log\frac{1}{v_\rho(S_{\varepsilon/2}(x))} \leqq K(\rho)^{\frac{1}{2}}\} \geqq 1-K(\rho)^{\frac{1}{2}},$$

or

$$\mu\{x \mid 1-v_\rho(S_{\varepsilon/2}(x)) \geqq 1-\exp\left[-K(\rho)^{\frac{1}{2}}\right]\} \geqq K(\rho)^{\frac{1}{2}}.$$

From (99)

$$m_1 \leqq 1-\exp\left[-K(\rho)^{\frac{1}{2}}\right]+K(\rho)^{\frac{1}{2}} \leqq 2K(\rho)^{\frac{1}{2}} \leqq 1-e^{-1}$$

and

$$1-m_1 \geqq 1-2K(\rho)^{\frac{1}{2}} \geqq e^{-1}.$$

Hence

$$(1-m_1)\log\frac{1}{1-m_1} \leqq \left[1-2K(\rho)^{\frac{1}{2}}\right]\log\frac{1}{1-2K(\rho)^{\frac{1}{2}}} \leqq 2K(\rho)^{\frac{1}{2}}.$$

Thus, (102) and Lemma 12 yield a universal constant D_1 such that

$$H_\varepsilon(X) \leqq \frac{K(\rho)}{\gamma}\log\frac{\gamma}{K(\rho)}+D_1K(\rho)+2(K(\rho))^{\frac{1}{2}},$$

which proves the lemma.

Note that D can be made arbitrarily close to 2 as $K(\rho) \to 0$.

We are now ready to state and prove Theorem 2, the main result of this paper.

THEOREM 2. *There exist universal constants C and D such that*

$$H_\varepsilon(X) \leqq I_\varepsilon(X)+\log^+ I_\varepsilon(X)+C, \tag{105}$$

and, for

$$I_\varepsilon(X) \leqq \tfrac{1}{8}(1-1/e)^2,$$

$$H_\varepsilon(X) \leqq D(I_\varepsilon(X))^{\frac{1}{2}}. \tag{106}$$

Furthermore, $I_\varepsilon(X) = I_\varepsilon'(X)$.

PROOF.

Choose, given $\eta > 0$,

$$\rho \in R_\varepsilon(X), \qquad \text{with } I_\varepsilon'(X) \leqq I(\rho) \leqq I_\varepsilon'(X)+\eta. \tag{107}$$

Then, by Lemma 13,

$$K(\rho) \leqq I'(X)+\eta. \tag{108}$$

By Lemma 14, on the other hand, we found

$$H_\varepsilon(X) \leqq K(\rho)+\log^+ K(\rho)+C,$$

so

$$H_\varepsilon(X) \leqq I_\varepsilon'(X)+\log^+[I_\varepsilon'(X)+\eta]+C+\eta,$$

for every $\eta > 0$. This proves (105) for $I_\varepsilon'(X)$ instead of $I_\varepsilon(X)$. By Lemma 6 and Lemma 3, if $I_\varepsilon'(X)$ is infinite, there is nothing to prove. If $I_\varepsilon(X)$ is finite, however, (109) forces $H_\varepsilon(X)$ to be finite. By Lemma 6, then,

$$I_\varepsilon(X) = I_\varepsilon'(X), \tag{110}$$

and (105) is proved in the desired form (if $I_\varepsilon'(X)$ is infinite, equality certainly holds).

Now let us prove (106). Let $I_\varepsilon'(X) \leqq \frac{1}{8}(1-1/e)^2$, and choose ρ by (107), with $\eta \leqq \frac{1}{8}(1-1/e)^2$. (108) holds, so that, by Lemma 15,

$$H_\varepsilon(X) \leqq D(I_\varepsilon'(X)+\eta)^{\frac{1}{2}}. \tag{111}$$

Since (111) holds for all sufficiently small $\eta > 0$, and since (110) holds, (106) is true. This completes the proof of the main theorem.

The rest of the paper is devoted to consequences of the main theorem.

8. Consequences of the main theorem. This final section gives various consequences of Theorem 2. The first result is of independent interest and also needed for the second result. We recall [7, page 1008] that $H_0(X)$ is defined as the infimum of the entropies of all partitions of X less a set of measure zero by atoms; when there is no such partition, $H_0(X)$ is infinite. With this definition, $H_\varepsilon(X)$ is continuous from above in ε, even at $\varepsilon = 0$, as was shown in the above cited reference.

COROLLARY 1. *$I_\varepsilon(X)$ is continuous from above in ε on $\varepsilon \geqq 0$.*

PROOF. Fix n a positive integer, and consider,

$$H_{\varepsilon'}(X^{(n)}) \leqq I_{\varepsilon'}(X^{(n)})+\log^+ I_{\varepsilon'}(X^{(n)})+C. \tag{112}$$

First let $H_\varepsilon(X) < \infty$. We can then choose an $\varepsilon' > \varepsilon$ so that

$$H_\varepsilon(X^{(n)}) < H_{\varepsilon'}(X^{(n)})+1, \tag{113}$$

since $H(X^{(n)})$ is continuous from above in ε.

Now

$$I_\varepsilon(X^{(n)}) = nI_\varepsilon(X)$$

from (7), so

$$\begin{aligned} nI_\varepsilon(X) &\leqq H_\varepsilon(X^{(n)}) < H_{\varepsilon'}(X^{(n)})+1 \\ &< nI_{\varepsilon'}(X)+\log^+ I_{\varepsilon'}(X)+\log n+C+1. \end{aligned}$$

We then find

$$I_\varepsilon(X) < I_{\varepsilon'}(X)+\frac{1}{n}\log^+ I_{\varepsilon'}(X)+(\log n+C+1)/n. \tag{114}$$

Since

$$I_{\varepsilon'}(X) \leqq I_{\varepsilon}(X) < \infty,$$

given $\eta > 0$, we can by (114) choose an n so large that

$$I_{\varepsilon}(X) < I_{\varepsilon'}(X)+\eta, \tag{115}$$

which proves continuity from above in ε in case $H_{\varepsilon}(X)$ is finite.

Now let $H_{\varepsilon}(X)$, and so $I_{\varepsilon}(X)$, be infinite. Given a large $N > 0$, we are to find an $\varepsilon_0 > \varepsilon$ such that

$$I_{\varepsilon'}(X) > N \quad \text{when} \quad \varepsilon < \varepsilon' < \varepsilon_0. \tag{116}$$

We can assume that

$$H_{\varepsilon'}(X) < \infty, \qquad \varepsilon' \geqq \varepsilon. \tag{117}$$

By the continuity of $H_{\varepsilon}(X)$ from above in ε, we can find an $\varepsilon_0(n) > \varepsilon$ such that

$$H_{\varepsilon'}(X^{(n)}) > n^2, \qquad \text{if} \quad \varepsilon < \varepsilon' \leqq \varepsilon_0(n). \tag{118}$$

Thus (112) becomes

$$n^2 < H_{\varepsilon'}(X^{(n)}) < nI_{\varepsilon'}(X)+\log^+ I_{\varepsilon'}(X)+\log n+C,$$

or

$$n < I_{\varepsilon'}(X)+\frac{1}{n}\log^+ I_{\varepsilon'}(X)+\frac{1}{n}(\log n+C) \quad \text{for} \quad \varepsilon < \varepsilon' \leqq \varepsilon_0(n). \tag{119}$$

But

$$\frac{1}{n}\log^+ I_{\varepsilon'}(X) \leqq \log^+ I_{\varepsilon'}(X) \leqq I_{\varepsilon'}(X),$$

so (119) can be written in the following form if $n \geqq n_0$, where $(\log n_0+C)/n_0 \leqq 1$:

$$I_{\varepsilon'}(X) \geqq \tfrac{1}{2}n-1 \quad \text{for} \quad \varepsilon < \varepsilon' \leqq \varepsilon_0(n). \tag{120}$$

Then (116) is proved by taking

$$n > \max(n_0, 2N+2).$$

Corollary 1 is proved.

The proof of Corollary 1 can be modified to prove continuity of $I_{\varepsilon}(X)$ in ε from below for certain cases in which it is known that $H_{\varepsilon}(X)$ is continuous in ε from below. One case is that of mean-continuous Gaussian processes on the unit interval, by a modification of [8b], Theorem 1, to prove that $H_{\varepsilon}(X^{(n)})$ is continuous from below in ε, $n \geqq 2$. Details are omitted.

COROLLARY 2. *For any probabilistic metric space X, one has*

$$H_{\varepsilon}(X) \sim I_{\varepsilon}(X) \quad \textit{as} \quad \varepsilon \to 0. \tag{121}$$

PROOF. Since

$$I_\varepsilon(X) \leqq H_\varepsilon(X) \leqq I_\varepsilon(X)+\log^+ I_\varepsilon(X)+C$$

by Theorem 2, the result is true in case $H_\varepsilon(X) \to \infty$ as $\varepsilon \to 0$. So suppose $H_\varepsilon(X)$ remains bounded as $\varepsilon \to 0$. Then by the continuity of $H_\varepsilon(X)$ from above in ε even at $\varepsilon = 0$ ([7], page 1008), we conclude that $H_0(X)$ is finite.

Now we always have

$$H_0(X) = I_0(X) \tag{122}$$

as (7) shows. By Corollary 1,

$$I_0(X) = \lim_{\varepsilon \to 0+} I_\varepsilon(X). \tag{123}$$

If $I_0(X) = 0$, then $H_0(X) = 0$, and, in fact,

$$H_\varepsilon(X) = I_\varepsilon(X) = 0, \qquad \text{all} \quad \varepsilon > 0,$$

and there is nothing to prove. If $I_0(X) > 0$, then $H_0(X) > 0$, and, by the continuity in ε from above of $H_\varepsilon(X)$

$$H_0(X) = \lim_{\varepsilon \to 0+} H_\varepsilon(X). \tag{124}$$

Equations (123) and (124) taken in conjunction with (122) give

$$\lim_{\varepsilon \to 0+} \frac{H_\varepsilon(X)}{I_\varepsilon(X)} = 1, \tag{125}$$

which proves (121) and Corollary 2.

The next batch of corollaries picks up some loose ends involving $I_\varepsilon'(X)$, defined in (19) as the inf of $I(\rho)$ for ρ in $R_\varepsilon(X)$. We restate the following, although it was proved in Theorem 2.

COROLLARY 3. $I_\varepsilon'(X) = I_\varepsilon(X)$.

COROLLARY 4. *Equation* (53) *holds for* $I_\varepsilon(X)$ *in place of* $I_\varepsilon'(X)$.

PROOF. Substitute $I_\varepsilon(X)$ for $I_\varepsilon'(X)$ in (53) to prove this corollary. We know of no other proof.

Now define for $0 \leqq \delta \leqq 1$

$$\bar{I}_{\varepsilon;\delta}(X) = \inf_{A \text{ measurable; } \mu(A) \leqq \delta} I_\varepsilon(X-A), \tag{126}$$

which is always finite for $\delta > 0$. Notice that

$$\bar{I}_{\varepsilon;0}(X) = I_\varepsilon(X).$$

Also,

$$\bar{I}_{\varepsilon;\delta}(X) \leqq H_{\varepsilon;\delta}(X). \tag{127}$$

The next lemma is needed in the proof of Corollary 5.

LEMMA 16. *Let X be a probabilistic metric space with $H_\varepsilon(X)$ finite. Then for every $\alpha > 0$ there exists a finite β such that, if Y is a measurable subset of X,*

$$\mu(Y) \geqq \alpha$$

implies

$$H_\varepsilon(Y) \leqq \beta.$$

PROOF. Let $U = \{U_i\}$ be an ε-partition of X with $\{\mu(U_i)\}$ non-increasing and

$$\sum \mu(U_i) \log \frac{1}{\mu(U_i)} = H_\varepsilon(X) < \infty.$$

Let

$$V = \{V_i\} = \{U_i \cap Y\},$$

an ε-partition of Y. Its entropy can be bounded as follows:

$$\begin{aligned} H(V) &= \sum \frac{\mu(V_i)}{\mu(Y)} \log \frac{\mu(Y)}{\mu(V_i)} \\ &= \frac{1}{\mu(Y)} \sum \mu(V_i) \log \frac{1}{\mu(V_i)} + \frac{1}{\mu(Y)} \log \mu(Y) \\ &\leqq \frac{1}{\alpha} \sum \mu(V_i) \log \frac{1}{\mu(V_i)} + \frac{1}{e}. \end{aligned}$$

Now $\mu(U_i) \leqq 1/e$, $i \geqq 2$, so

$$\begin{aligned} H(V) &\leqq \frac{1}{\alpha}\left(\frac{2}{e} + \sum_{i \geq 2} \mu(U_i) \log \frac{1}{\mu(U_i)}\right) + \frac{1}{e} \\ &\leqq \frac{1}{\alpha}\left(\frac{2}{e} + H_\varepsilon(X)\right) + \frac{1}{e}. \end{aligned}$$

Thus, if we define β by

$$\beta = \frac{1}{\alpha}\left(\frac{2}{e} + H_\varepsilon(X)\right) + \frac{1}{e},$$

then

$$H_\varepsilon(Y) \leqq H(V) \leqq \beta,$$

and Lemma 16 is proved.

We then have the following result, proved for $H_\varepsilon(X)$ in [7], Theorem 5.

COROLLARY 5. *$I_\varepsilon(X) = \lim_{\delta \to 0+} \bar{I}_{\varepsilon;\delta}(X)$, and, in fact, $\bar{I}_{\varepsilon;\delta}(X)$ is continuous from above in δ, $1 > \delta \geqq 0$. Also, given $\eta > 0$, there is a $\delta_0 > 0$ such that, for A measurable of measure at most δ_0,*

$$|I_\varepsilon(X - A) - I_\varepsilon(X)| < \eta, \tag{128}$$

if $I_\varepsilon(X)$ is finite. If $I_\varepsilon(X)$ is infinite, then, given $N > 0$, there is a $\delta_0 > 0$ such that, for A measurable of measure at most δ_0,

$$I_\varepsilon(X-A) > N. \tag{129}$$

Finally, if $\delta > 0$, $\bar{I}_{\varepsilon;\delta}(X)$ is continuous from below in δ if X is nonatomic.

PROOF. We first prove (128). Let n be so large that

$$\left|\frac{1}{n} H_\varepsilon(X^{(n)}) - I_\varepsilon(X)\right| < \frac{\eta}{3}. \tag{130}$$

Let δ be so small that $\delta \leqq \frac{1}{2}$ and

$$\left|\frac{1}{n} H_\varepsilon(X^{(n)}-B) - \frac{1}{n} H_\varepsilon(X^{(n)})\right| < \frac{\eta}{3}, \tag{131}$$

provided $\mu^{(n)}(B) \leqq \delta$. Then

$$\left|\frac{1}{n} H_\varepsilon(X^{(n)}-B) - I_\varepsilon(X)\right| \leqq \frac{2\eta}{3},$$

whenever $\mu^{(n)}(B) \leqq \delta$.

Define δ_0 such that

$$(1-\delta_0)^n = 1-\delta: \tag{132}$$

$\delta_0 \leqq \frac{1}{2}$. If A is measurable with $\mu(A) \leqq \delta_0$, then

$$\mu^{(n)}((X-A)^{(n)}) \geqq 1-\delta, \qquad \mu^{(n)}(X^{(n)} - (X-A)^{(n)}) \leqq \delta,$$

and

$$\left|\frac{1}{n} H_\varepsilon((X-A)^{(n)}) - I_\varepsilon(X)\right| \leqq \frac{2\eta}{3}: \tag{133}$$

Now by Theorem 2,

$$\left|H_\varepsilon(X-A)^{(n)}) - I_\varepsilon((X-A)^{(n)})\right| \leqq \log^+ I_\varepsilon((X-A)^{(n)}) + C,$$

or

$$\left|\frac{1}{n} H_\varepsilon((X-A)^{(n)}) - I_\varepsilon((X-A)\right| \leqq \frac{1}{n} \log^+ I_\varepsilon(X-A) + \frac{1}{n}(\log n + C). \tag{134}$$

By Lemma 16,

$$\sup_{\mu(A)\leqq\frac{1}{2}} H_\varepsilon(X-A) \leqq E \quad \text{say.} \tag{135}$$

Choose n in advance so large that

$$\frac{1}{n} \log^+ E + \frac{1}{n}(\log n + C) < \frac{\eta}{3}. \tag{136}$$

Then (134) becomes

$$\left|\frac{1}{n} H_\varepsilon((X-A)^{(n)}) - I_\varepsilon(X-A)\right| < \frac{\eta}{3}, \tag{137}$$

and (133) then yields (128).

Now to prove (129). In this case, (131) becomes

$$\frac{1}{n} H_\varepsilon(X^{(n)}-B) > 2N \tag{138}$$

whenever

$$\mu^{(n)}(B) \leqq \delta.$$

Define δ_0 by (132), so that (133) becomes

$$\frac{1}{n} H_\varepsilon((X-A^{(n)}) > 2N \tag{139}$$

if $\mu(A) \leqq \delta_0$.

For $H_\varepsilon(X-A)$ finite, the only case of interest, (134) becomes

$$\frac{1}{n} H_\varepsilon((X-A)^{(n)}) \leqq I_\varepsilon(X-A) + \frac{1}{n}\log^+ I_\varepsilon(X-A) + \frac{1}{n}(\log n + C). \tag{140}$$

If

$$\frac{1}{n}\log^+ I_\varepsilon(X-A) \geqq \frac{1}{n}\log N,$$

then (129) holds. So assume

$$\frac{1}{n}\log^+ I_\varepsilon(X-A) \leqq \frac{1}{n}\log N. \tag{141}$$

Given N in advance, we choose n in advance so that

$$\frac{1}{n}(\log N + \log n + C) < N,$$

which makes (140) and (139) imply

$$2N < \frac{1}{n} H_\varepsilon((X-A)^{(n)}) < I_\varepsilon(X-A) + N,$$

which proves (129).

To prove the continuity of $\bar{I}_{\varepsilon;\delta}(X)$ in δ from above, we can assume $\delta > 0$; the case $\delta = 0$ is done by (128) and (129). Given $\bar{I}_{\varepsilon;\delta}(X)$, finite since $\delta > 0$, and $\eta > 0$, we are to prove that there is a $\lambda_0 > 0$ with

$$I_\varepsilon(X-A) \geqq I_{\varepsilon;\delta}(X) - \eta \tag{143}$$

whenever $\mu(A) \leqq \delta+\lambda_0$. And if $\mu(A) \leqq \delta$, there is nothing to prove, since then (143) is satisfied even with $\eta = 0$.

Let us digress to prove the following result:
Given $\delta > 0$, there is a $\lambda > 0$ such that if A satisfies $\mu(A) \geqq \delta$, then, for some x in A,

$$\mu(S_{\varepsilon/2}(x) \cap A) \geqq \lambda. \tag{144}$$

Here is a proof of that fact. Let B be such that $\mu(B) \geqq 1-\delta/2$ and $H_\varepsilon(B) < \infty$. Consider $B \cap A$, which has

$$\mu(B \cap A) \geqq \delta/2.$$

By Lemma 16,

$$H_\varepsilon(B \cap A) < E$$

for some constant E. By Lemma 8, there is an x in $B \cap A$ with

$$\frac{2}{\delta}\mu(S_{\varepsilon/2}(x) \cap B \cap A) \geqq e^{-E},$$

and so, for some x in A, (144) holds, with

$$\lambda = \frac{\delta}{2} e^{-E}.$$

To prove (143), let

$$\delta \leqq \mu(A) \leqq \delta+\lambda.$$

Write $W = S_{\varepsilon/2}(x)$ where $\mu(W \cap A) \geqq \lambda$. As in [7], Theorem 1, we can reduce to the case X nonatomic, in which case, given $\lambda_0 \leqq \lambda$, there is a subset of V of $W \cap A$ with $\mu(V) = \lambda_0$. Consider the set $(X-A) \cap V$, where

$$\mu((X-A) \cup V) \geqq 1-\delta.$$

This forces

$$I_\varepsilon((X-A) \cup V) \geqq \bar{I}_{\varepsilon;\delta}(X). \tag{145}$$

On the other hand, Corollary 4 gives

$$I_\varepsilon((X-A) \cup V) \leqq pI_\varepsilon(X-A)+(1-p)I_\varepsilon(V)+H(p), \tag{146}$$

with

$$p = \mu(X-A)/(\mu(X-A)+\mu(V));$$

$$H(p) = p \log \frac{1}{p}+(1-p) \log \frac{1}{1-p}.$$

Since

$$V \subset D_{\varepsilon/2}(x),$$

however,

$$I_\varepsilon(V) = 0,$$

and (146) becomes

$$I_\varepsilon((X-A)\cup V) \leqq pI_\varepsilon(X-A)+H(p). \tag{147}$$

Together with (145), (147) gives

$$I_\varepsilon(X-A) \leqq \frac{1}{p}\bar{I}_{\varepsilon;\delta}(X)-\log\frac{1}{p}-\frac{1-p}{p}\log(1-p). \tag{148}$$

Now

$$1<\frac{1}{p}\leqq 1+\frac{\lambda_0}{1-\delta}. \tag{149}$$

If we choose λ_0 so small that

$$\log\left(1+\frac{\lambda_0}{1-\delta}\right)+\frac{\lambda_0}{1-\delta}\log\frac{1}{H\left(\frac{1-\delta}{\lambda_0}\right)}<\eta,$$

(143), and hence continuity from above, follows. To prove continuity from below when X is nonatomic is similar and omitted. This completes the proof of Corollary 5.

To get a stronger "strong converse" define, as in Lemma 5,

$$I_\varepsilon''(X) = \inf_{[\delta_n]\to 0}\liminf_{n\to\infty}\frac{1}{n}H_{\varepsilon;\delta_n}(X^{(n)}), \tag{150}$$

so that

$$I_\varepsilon''(X) = I_\varepsilon(X) \tag{151}$$

if $H_\varepsilon(X)<\infty$. As in Lemma 5, $I_\varepsilon''(X)$ is the amount of capacity needed to transmit outcomes of X when it is desired that an arbitrarily large fraction of a block of outcomes be known to within ε, with probability approaching 1. Alternatively, $I_\varepsilon''(X)$ is the Rate Distortion Function for zero distortion. The stronger converse is really just that $I_\varepsilon(X)=I_\varepsilon''(X)$. This means that if K is a memoryless channel of capacity $C<I_\varepsilon(X)$, we cannot transmit outcomes of X over K such that, with probability approaching 1, an arbitrarily large fraction of a long block of outcomes is known to within ε.

First recall

$$I_{\varepsilon;[\delta]}(X) = \liminf_{n\to\infty}\frac{1}{n}H_{\varepsilon;\delta}(X^{(n)})$$

for $0\leqq\delta<1$. We then have the following corollary.

COROLLARY 6.

$$I_{\varepsilon;[\delta]}(X) = \liminf_{n\to\infty}\frac{1}{n}\bar{I}_{\varepsilon;\delta}(X^{(n)}).$$

PROOF. If

$$H_{\varepsilon;\delta}(X^{(n)}) \to \infty \qquad \text{as} \quad n \to \infty,$$

then (105) shows

$$H_{\varepsilon;\delta}(X^{(n)}) \sim \bar{I}_{\varepsilon;\delta}(X^{(n)}).$$

If, on the other hand,

$$H_{\varepsilon;\delta}(X^{(n)}) \leqq F, \qquad \text{say,} \tag{152}$$

for all $\eta > 0$, then we show that $H_\varepsilon(X) = 0$, and so

$$I_\varepsilon(X) = \bar{I}_{\varepsilon;\delta}(X) = H_{\varepsilon;\delta}(X^{(n)}) = \bar{I}_{\varepsilon;\delta}(X^{(n)}) = 0,$$

and the result follows. So suppose (152) holds. Then, by Lemma 8, there is a constant $E > 0$ such that, for each $n \geqq 1$, there is an ε-set A_n in $X^{(n)}$ with

$$\mu^{(n)}(A_n) \geqq E.$$

On the other hand,

$$\mu^{(n)}(A_n) \leqq [\max_{S \text{ an } \varepsilon\text{-set}} \mu(S)]^n,$$

so

$$\max_{S \text{ an } \varepsilon\text{-set}} \mu(S) = 1,$$

and $H_\varepsilon(X) = 0$, which proves Corollary 6.

In order to prove the stronger converse, and strengthen Lemmas 4 and 5 by removing the requirement from them that $H_\varepsilon(X)$ be finite, we need a lemma. We first need to define the class $R_{\varepsilon;\eta}(X)$ of probability distributions for $0 \leqq \eta$; note that

$$R_{\varepsilon;0}(X) = R_\varepsilon(X).$$

DEFINITION. ${}_RR_{\varepsilon;\eta}(X)$ is that class of Borel measures ρ on $X \times X$ whose marginal distribution on the first coordinate is μ, and such that

$$\rho(\{(x, y) \mid d(x, y) \leqq \varepsilon/2\}) \geqq 1 - \eta.$$

The class ${}_DR_{\varepsilon;\eta}(X)$ is defined similarly for the diametric case.

Then define

$$I_{\varepsilon;\eta}(X) = \inf_{\rho \in R_{\varepsilon;\eta}(X)} I(\rho),$$

so that $I_{\varepsilon;\eta}(X)$ is the Rate Distortion function evaluated for distortion η. Therefore,

$$I_{\varepsilon;\eta}(X) \leqq \bar{I}_{\varepsilon;\eta}(X) + \eta \log \frac{1}{\eta},$$

and

$$I_{\varepsilon;0}(X) = I_\varepsilon(X).$$

Finally, define

$$I_\varepsilon'''(X) = \lim_{\eta \to 0+} I_{\varepsilon;\eta}(X).$$

We have been unable to determine, except when either is 0 or ∞, in which case the answer is affirmative, whether

$$I_\varepsilon'''(X) = I_\varepsilon(X).$$

The next lemma shows that the two are however asymptotic when either is large. A strengthening of Theorem 2, it is enough for the stronger converse. The result for $\eta > 0$ gives a result on the relation between $H_{\varepsilon;\eta^{\frac{1}{2}}}(X)$ and $I_{\varepsilon;\eta}(X)$.

LEMMA 17. *Let* $0 \leqq \eta \leqq 1/e^2$. *Define*

$$L_{\varepsilon;\eta}(X) = \frac{1}{1-\eta}\left(I_{\varepsilon;\eta}(X)+\frac{2}{e}\right)+\log\frac{1}{1-\eta}.$$

Then, for C *the same constant as in Theorem* 2,

$$H_{\varepsilon;\eta^{\frac{1}{2}}}(X < (1+6\eta^{\frac{1}{2}})\{L_{\varepsilon;\eta}(X)+\log^+ L_{\varepsilon;\eta}(X)+C\}+3\eta^{\frac{1}{2}}: \tag{153}$$

$$H_\varepsilon(X) \leqq I_\varepsilon'''(X)+\log^+ I_\varepsilon'''(X)+C+\tfrac{4}{3}. \tag{154}$$

PROOF. Given $\lambda > 0$, choose ρ in $R_{\varepsilon;\eta}(X)$ with

$$I(\rho) \leqq I_{\varepsilon;\eta}(X)+\lambda. \tag{155}$$

Let S be that subset of $X \times X$ on which, in the notation of Lemma 14,

$$A_x \subset S_{\varepsilon/2}(x),$$

so that

$$\rho(S) \geqq 1-\eta.$$

Define ρ_1 to be the probability measure on $X \times X$ given by

$$\rho_1(T) = \rho(S \cap T)/\rho(S),$$

with marginals μ_1 and ν_1 respectively. We observe, but do not use, that

$$\frac{d\mu_1(x)}{d\mu(x)} = \frac{1}{\rho(S)}\int_{S_x}\frac{d\rho(x, y)}{d\mu(x)d\nu_\rho(y)}\,d\nu_\rho(y),$$

$$\frac{d\nu_1(y)}{d\nu_\rho(y)} = \frac{1}{\rho(S)}\int_{S_y}\frac{d\rho(x, y)}{d\mu(x)d\nu_\rho(y)}\,d\mu(x).$$

More important for our purpose is

$$\frac{d\rho_1(x, y)}{d\rho(x, y)} = 0, \qquad (x, y) \in S^c, \tag{156}$$

$$= \frac{1}{\rho(S)}, \qquad (x, y) \in S.$$

As a consequence of (156), we have

$$\frac{d\rho_1(x, y)}{d\mu_1(x)d\nu_1(y)} = 0, \qquad (x, y) \in S^c \tag{157}$$

$$= \frac{d\rho(x, y)}{d\mu(x)d\nu_\rho(y)}\,\frac{1}{\rho(S)}\,\frac{d\mu(x)}{d\mu_1(x)}\,\frac{d\nu_\rho(y)}{d\nu_1(y)}, \qquad (x, y) \in S.$$

From (157), we can conclude

$$I(\rho_1) = \frac{1}{\rho(S)}\int_S \log\left(\frac{d\rho}{d\mu dv_\rho}\right) d\rho + \log\frac{1}{\rho(S)} + \int \log\left(\frac{d\mu}{d\mu_1}\right) d\rho_1 + \int \log\left(\frac{dv_\rho}{dv_1}\right) d\rho_1. \tag{158}$$

Continuing from (158), we have the inequality

$$I(\rho_1) \leqq \frac{1}{\rho(S)}\int \left|\log\frac{d\rho}{d\mu dv_\rho}\right| d\rho + \log\frac{1}{\rho(S)} + \int\left(\log\frac{d\mu}{d\mu_1}\right)\frac{d\mu_1}{d\mu}\, du + \int\left(\log\frac{dv_\rho}{dv_1}\right)\frac{dv_1}{dv_\rho}\, dv_\rho$$

which yields

$$I(\rho_1) \leqq \frac{1}{\rho(S)}\left(I(\rho) + \frac{2}{e}\right) + \log\frac{1}{\rho(S)}. \tag{159}$$

The first term in (159) arises from the fact that the function $t \log t \geqq -1/e$ in $t \geqq 0$; there is no other term because $-t \log t$ is convex in $t \geqq 0$, and

$$\int \frac{d\mu_1}{d\mu}\, d\mu = \int \frac{dv_1}{dv_\rho}\, dv_\rho = 1.$$

Now form the random partition of Lemma 14, using the measure ρ_1, instead of ρ. Introduce the notation

$$H(U; \mu)$$

for U an ε-partition of X and μ a Borel probability measure on X to denote the entropy of U when the measure in question is μ. From Lemma 14, we conclude that there is an ε-partition U of a set of μ_1-measure 1 in X such that

$$H(U; \mu_1) \leqq I(\rho_1) + \log^+ I(\rho_1) + C, \tag{160}$$

for the C of Theorem 2. However, there are several obstacles to translating (160) into an inequality that will lead to (153).

The first of these is that U is an ε-partition of X with respect to μ_1. What can be left uncovered by U? If Z denotes the set of X not covered by U, then observe that

$$Z \times X \subset S^c.$$

Now let

$$U = \{U_i\}, \qquad i \geqq 1:$$

we have

$$\mu\left(\bigcup_i U_i\right) = \lambda \geqq \rho(S) \geqq 1 - \eta.$$

Let F be the set of indices i such that

$$\rho(U_i \times X)/\rho((U_i \times X) \cap S) \geqq 1 + 2\eta^{\frac{1}{2}} \tag{161}$$

for i in F. If we define

$$V_F = \bigcup_{i \in F} U_i,$$

then we observe that V_F and Z are disjoint. Furthermore, we claim that

$$\rho(V_F \times X) = \mu(V_F) \leqq \eta^{\frac{1}{2}} - \mu(Z). \tag{162}$$

For if not, then, since

$$\rho(V_F \times X)/\rho((V_F \times X) \cap S) \geqq 1 + 2\eta^{\frac{1}{2}},$$

we have

$$\rho((V_F \times X) \cap S) \leqq \rho(V_F \times X)/(1 + 2\eta^{\frac{1}{2}}),$$

or

$$\rho(V_F \times X) - \rho((V_F \times X) \cap S) \geqq \frac{2\eta^{\frac{1}{2}}}{1 + 2\eta^{\frac{1}{2}}} \rho(V_F \times X).$$

Consequently,

$$\rho(V_F \times X) - \rho((V_F \times X) \cap S) \geqq \frac{2\eta}{1 + 2\eta^{\frac{1}{2}}} - \frac{2\eta^{\frac{1}{2}}}{1 + 2\eta^{\frac{1}{2}}} \mu(Z) \geqq \eta - \mu(Z),$$

since $\eta < \frac{1}{4}$, we would therefore have

$$\rho((V_F \times X) \cap S^c) + \rho((Z \times X) \cap S^c) > \eta,$$

and so

$$\rho(S^c) > \eta,$$

a contradiction. We conclude

$$\mu(V_F \cup Z) \leqq \eta^{\frac{1}{2}}.$$

Now let U' be the $\varepsilon; \eta^{\frac{1}{2}}$-partition of X with respect to μ consisting of those sets U_i of U with i not in F. The second obstacle to overcome is to estimate $H(U'; \mu)$ in terms of $H(U; \mu)$.

We have

$$\begin{aligned} H(U' : \mu) &= \sum_{i \notin F} \frac{\mu(U_i)}{\lambda - \mu(V_F)} \log \frac{\lambda - \mu(V_F)}{\mu(U_i)} \\ &\leqq \frac{1}{\lambda - \mu(V_F)} \sum_{i \notin F} \mu(U_i) \log \frac{1}{\mu(U_i)} \qquad (163) \\ &= \frac{1}{\lambda - \mu(V_F)} T, \qquad \text{say.} \end{aligned}$$

From the convexity of the function $t \log (1/t)$ in $t \geqq 0$, however, we can conclude

$$T \leqq (\lambda - \mu(V_F)) \log \frac{1 - \mu_1(V_F)}{\lambda - \mu(V_F)} + \sum_{i \notin F} \mu(U_i) \log \frac{1}{\mu_1(U_i)},$$

and so

$$T \leqq ((\lambda - \mu(V_F)) \log \frac{1 - \mu_1(V_F)}{\lambda - \mu(V_F)} + \sum_{i \notin F} \rho(U_i \times X) \log \frac{1}{\rho((U_i \times X) \cap S}. \tag{164}$$

(164), coupled with the definition (161) of F, yields

$$T \leqq (\lambda-\mu(V_F)) \log \frac{1-\mu_1(V_F)}{\lambda-\mu(V_F)}$$

$$+(1+2\eta^{\frac{1}{2}}) \sum_{\text{all } i} \rho((U_i \times X) \cap S) \log \frac{1}{\rho((U_i \times X) \cap S)},$$

and finally

$$T \leqq (\lambda-\mu(V_F)) \log \frac{1}{\lambda-\mu(V_F)} + (1+2\eta^{\frac{1}{2}}) H(U: \mu_1). \tag{165}$$

(163) coupled with (165) now gives

$$H_{\varepsilon;\eta^{\frac{1}{2}}}(X) < \log \frac{1}{\lambda-\mu(V_F)} + \frac{1+2\eta^{\frac{1}{2}}}{\lambda-\mu(V_F)} H(U: \mu_1), \tag{166}$$

and finally

$$H_{\varepsilon;\eta^{\frac{1}{2}}}(X) < 3\eta^{\frac{1}{2}} + (1+6\eta^{\frac{1}{2}}) H(U: \mu_1). \tag{167}$$

Combining (167) with (160), (159), and (155) proves

$$H_{\varepsilon;\eta^{\frac{1}{2}}}(X) < 3\eta^{\frac{1}{2}} + (1+6\eta^{\frac{1}{2}}) \left\{ \frac{1}{1-\eta} \left(I_{\varepsilon;\eta}(X) + \frac{2}{e} \right) + \log \frac{1}{1-\eta} \right.$$

$$\left. + \log^+ \left[\frac{1}{1-\eta} \left(I_{\varepsilon;\eta}(X) + \frac{2}{e} \right) + \log \frac{1}{1-\eta} \right] + C \right\}, \tag{168}$$

which thus proves (153). To prove (154), let $\eta \to 0^+$ in (168). We know from [5], Theorem 4, that

$$\lim_{\delta \to 0^+} H_{\varepsilon;\delta}(X) = H_\varepsilon(X),$$

and

$$\lim_{\delta \to 0^+} I_{\varepsilon;\delta}(X) = I_\varepsilon'''(X),$$

by definition. The simple inequality

$$\log \left(1 + \frac{2}{e} \right) + \frac{2}{e} < \frac{4}{3}$$

completes the proof of (154) and hence of Lemma 17.

Now comes the stronger converse; define as in Lemma 5

$$I_\varepsilon''(X) = \inf_{[\delta_n] \to 0} \liminf_{n \to \infty} \frac{1}{n} H_{\varepsilon;\delta_n}(X^{(n)}).$$

Lemma 5 proved Corollary 7 when $H_\varepsilon(X)$ is finite.

COROLLARY 7. $I_\varepsilon''(X) = I_\varepsilon(X)$.

PROOF. Since

$$I_\varepsilon''(X) \leqq I_\varepsilon(X),$$

there is nothing to prove if $I_\varepsilon''(X)$ is infinite. So let $I_\varepsilon''(X)$ be finite. Then, as in Lemma 6, given an $\varepsilon;\delta_n$ partition U of $X^{(n)}$, such that

$$\frac{1}{n}H(U) \leqq I_\varepsilon''(X)+\eta, \qquad \text{say,} \qquad \text{with } \eta > 0,$$

we can obtain a ρ in $R_{\varepsilon;\delta_n}(X)$ with

$$I(\rho) \leqq (1-\delta_n)\left[I_\varepsilon''(X)+\eta+\log\frac{1}{1-\delta_n}\right]+\delta_n \log\frac{1}{\delta_n}.$$

That is,

$$I_{\varepsilon;\delta_n}(X) \leqq (1-\delta_n)\left[I_\varepsilon''(X)+\eta+\log\frac{1}{1-\delta_n}\right]+\delta_n \log\frac{1}{\delta_n}.$$

By the definition of $I_\varepsilon''(X)$, then,

$$I_\varepsilon'''(X) \leqq I_\varepsilon''(X)+\eta.$$

Hence $I_\varepsilon'''(X)$ is finite, and by Lemma 17, so is $H_\varepsilon(X)$. By Lemma 5, then,

$$I_\varepsilon''(X) = I_\varepsilon(X),$$

as promised. Corollary 7 is proved, and the restriction that $H_\varepsilon(X)$ be finite is removed from Lemma 5.

We can likewise show the following, which removes the assumption from Lemma 4 that $H_\varepsilon(X)$ be finite.

COROLLARY 8. $I_\varepsilon(X) = \lim_{\delta\to 0^+} I_{\varepsilon;[\delta]}(X)$.

PROOF. As in Corollary 7, we construct for every $\delta > 0$, $\eta > 0$, a ρ in $R_{\varepsilon;\delta}(X)$ such that

$$I(\rho) \leqq (1-\delta)\left[I_{\varepsilon;[\delta]}(X)+\eta+\log\frac{1}{1-\delta}\right]+\delta \log\frac{1}{\delta}.$$

Let the limit in the statement of the corollary be finite and equal to I say. Then we find that $I_\varepsilon'''(X)$ is also bounded by I. That is, $I_\varepsilon'''(X)$, and so $H_\varepsilon(X)$, are finite, whenever

$$\lim_{\delta\to 0^+} I_{\varepsilon;[\delta]}(X) \tag{169}$$

is finite. By Lemma 4, then, we can assume (169) infinite. But since the limit in (169) is at most $I_\varepsilon(X)$, $I_\varepsilon(X)$ is infinite if (160) is infinite. This proves Corollary 8.

We now summarize a combination of our principal results. If K is a memoryless channel with capacity Γ less than $H_\varepsilon(X)-\log^+ H_\varepsilon(X)-C$, C a universal constant (or of finite capacity if $H_\varepsilon(X)$ is infinite) then it is not possible to transmit outcomes of X over K such that, with probability approaching 1, an arbitrarily large fraction

of a sequence of outcomes are known within ε or even a little more than ε. On the other hand, if Γ exceeds $H_\varepsilon(X)$, then outcomes of X can be transmitted over K such that, with probability approaching 1, all of a sequence of outcomes are known within ε, and such that block coding of outcomes need not be done before the channel encoder. In other words, the "one-shot" epsilon entropy tells whether the maximum error in a long block can be kept as little more than ε as we please, with probability approaching 1.

We close the paper with the outstanding open problem in this theory. Can

$$H_\varepsilon(X)-I_\varepsilon(X)$$

be arbitrarily large when either is finite? More generally, determine the function

$$f(x) = \sup_{H\varepsilon(X)\leq x}[H_\varepsilon(X)-I_\varepsilon(X)],$$

which is finite-valued if x is finite. It is clear that $f(x)$ is strictly increasing as x increases, that $f(0) = 0$, and that $f(x)$ is continuous in $x \geqq 0$. The main problem is whether $f(x)$ is bounded in $x \geqq 0$.

REFERENCES

[1] FANO, R. M. (1961). *Theory of Information.* Wiley, New York.

[2] KOLMOGOROV, ANDREI N. (1956). On the Shannon theory of information transmission in the case of continuous signals. *IEEE Trans. Information Theory* **IT-2** 102–108.

[3] MCELIECE, ROBERT J. and POSNER, EDWARD C. (1971). Hide and seek, data storage, and entropy. *Ann. Math. Statist.* **42** 1706–1716.

[4] PINKSTON, JOHN T. III (1967). Encoding independent sample information sources. Res. Lab. Elect. Technical Report No. 462, Massachusetts Institute of Technology.

[5] PINSKER, M. S. (1964). *Information and Information Stability of Random Variables and Processes.* (Translated from Russian.) Holden-Day, San Francisco.

[6] POSNER, EDWARD C. and RODEMICH, EUGENE R. (1969). Differential entropy and tiling. *J. Statist. Phys.* **1** 57–69.

[7] POSNER, EDWARD C., RODEMICH, EUGENE E. and RUMSEY, HOWARD JR. (1967). Epsilon entropy of stochastic processes. *Ann. Math. Statist.* **38** 1000–1020.

[8*a*] POSNER, EDWARD C., RODEMICH, EUGENE R. and RUMSEY, HOWARD JR. (1969). Product entropy of Guassian distributions. *Ann. Math. Statist.* **40** 870–904.

[8*b*] POSNER, EDWARD C., RODEMICH, EUGENE R. and RUMSEY, HOWARD JR. (1969). Epsilon entropy of Gaussian processes. *Ann. Math. Statist.* **40** 1272–1296.

[9] PROKHOROV, YU V. (1956). Convergence of random processes and limit theorems in probability theory. (English translation in *Theor. Probability Appl.* **1** 157–214.)

[10] SAKRISON, DAVID J. (1968). The rate distortion function for a class of sources. Space Sciences Laboratory Report, Series No. 9, Univ. of California, Berkeley.

[11] SHANNON, CLAUDE E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **17** 623–656.

[12] SHANNON, CLAUDE E. (1960). Coding theorems for a discrete source with a fidelity criterion. *Information and Decision Processes.* 93–126. McGraw-Hill, New York.

[13] VITUSHKIN, A. (1962). *Complexity of the Tabulating Problem.* (Translated from Russian as *Theory of the Transmission and Processing of Information.*) Pergamon, New York.