ON THE CODING THEOREM FOR THE NOISELESS CHANNEL[1]

By Patrick Billingsley

University of Chicago

1. Introduction. The purpose of this paper is to examine the coding theorem for a noiseless channel from a point of view different from the usual one. The idea is to take the base s expansion of a point in the unit interval as a realization of the stochastic process to be coded, and then to relate the compression a given coding achieves to known properties of the unit interval, properties connected with Hausdorff dimension and the Shannon-McMillan theorem. This leads to results which in certain ways are sharper than the ones previously obtained.

Let $\Omega = (0, 1]$ and let $\mathcal{B}$ consist of the Borel subsets of Ω. With each ω we associate its nonterminating base s expansion: $\omega = \sum_{n=1}^{\infty} x_n(\omega)/s^n$, where $x_n(\omega) = 0, 1, \cdots, s - 1$. Then each x_n is a measurable function on Ω. If μ is a probability measure on $\mathcal{B}$ then $\{x_1, x_2, \cdots\}$ becomes a stochastic process. Moreover, any stochastic process with state space (or alphabet)

$$\sigma = \{0, 1, \cdots, s - 1\}$$

can be represented in this form, provided it is atomless. More precisely, let $\{p(a_1, \cdots, a_n)\}$ be any consistent set of finite-dimensional distributions with the property that

$$\lim_{n\to\infty} p(a_1, \cdots, a_n) = 0$$

for any sequence $(a_1, a_2, \cdots)$ of elements of σ. Then there exists a measure μ on $\mathcal{B}$ such that

$$\mu\{\omega : x_k(\omega) = a_k, k = 1, \cdots, n\} = p(a_1, \cdots, a_n).$$

Clearly μ will be atomless, or continuous. We will be concerned with such atomless measures μ under which the process $\{x_n\}$ is stationary and ergodic, that is, with measures μ such that if T is defined by $T\omega = [s\omega]$ then T preserves μ and is ergodic under μ. This representation of a process has been used for other purposes by Harris [7].

For the purposes of this paper a *code is a continuous, nondecreasing function* ϕ *on* $[0, 1]$ *with* $\phi(0) = 0$ *and* $\phi(1) = 1$. With each ω we associate the nonterminating base s expansion of $\phi(\omega) : \phi(\omega) = \sum_{n=1}^{\infty} y_n(\omega)/s^n$, where

$$y_n(\omega) = 0, 1, \cdots, s - 1.$$

Thus ϕ is a scheme for associating with each sequence $x = (x_1, x_2, \cdots)$ of symbols from σ another such sequence $y = (y_1, y_2, \cdots)$. (For simplicity of

Received May 17, 1960.

[1] Research carried out at the Statistical Research Center, University of Chicago, under partial sponsorship of the Statistics Branch, Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org

notation we consider only codes with the same input and output alphabets.) The code ϕ has the desirable property that for any atomless probability measure on $\mathcal{B}$, there is probability one that in order to determine the first n elements of y one need only know a finite number of elements of x. A second desirable property would be that x can be uniquely recovered from y. We will, for each probability measure on $\mathcal{B}$, produce a code for which this recoverability condition holds, with probability one, and which is optimal, in a certain sense, among all codes, even those not having this property.

Note that if ϕ is simply assumed to be a mapping from [0, 1] to [0, 1] such that for *all* x and y, x is uniquely recoverable from y and the first n elements of y are determined by some finite number of elements of x, then it follows that ϕ is continuous and either increasing or decreasing. The definition above constitutes a slight weakening of these requirements—there is no real loss of generality in excluding the decreasing case and requiring $\phi(0) = 0, \phi(1) = 1$.

The efficiency of a code is measured by the amount it compresses a sequence $(x_1, \cdots, x_n)$. For any $\omega \in \Omega$ and $n \geqq 1$, let

$$u_n(\omega) = \{\omega' : x_k(\omega') = x_k(\omega), k = 1, \cdots, n\}.$$

Thus $u_n(\omega)$ is that s-adic interval of rank n, that is, that interval of the form $(l/s^n, (l + 1)/s^n]$, which contains ω. Now if the first n symbols of the expansion of ω are known then $u = u_n(\omega)$ is known, and it is known that $\phi(\omega) \in \phi(u)$. If $\langle \phi(u) \rangle$ denotes the smallest s-adic interval containing $\phi(u)$, then the number of symbols in the expansion of $\phi(\omega)$ which can be determined at this stage is exactly the rank of the s-adic interval $\langle \phi(u) \rangle$. But the rank of $\langle \phi(u) \rangle$ is clearly $-\lg_s \lambda\langle \phi(u) \rangle$, where λ denotes Lebesgue measure. Thus the first n symbols in the expansion of ω determine exactly $-\lg_s \lambda\langle \phi(u_n(\omega)) \rangle$ symbols in the expansion of $\phi(\omega)$. Therefore the compression effected by the code ϕ on the first n symbols in the expansion of ω is

$$C_n(\omega) = -n^{-1} \lg_s \lambda\langle \phi(u_n(\omega)) \rangle. \tag{1.1}$$

To simplify the mathematics we will, in the first part of the paper, replace $C_n(\omega)$ by

$$D_n(\omega) = -n^{-1} \lg_s \lambda(\phi(u_n(\omega))). \tag{1.2}$$

(See the remarks at the end of the following paragraph.)

A code is efficient if $C_n(\omega)$ is small in some asymptotic sense. Let

$$C_\phi(\omega) = \lim_{n\to\infty} C_n(\omega),$$

if this limit exists, and let

$$\overline{C}_\phi(\omega) = \liminf_{n\to\infty} C_n(\omega).$$

Define $D_\phi(\omega)$ and $\overline{D}_\phi(\omega)$ similarly in terms of $D_n(\omega)$. Suppose we are given a stationary, atomless, ergodic probability measure μ. If $F(\alpha) = \mu(0, \alpha]$, that is if F is the distribution function corresponding to μ, then F is a code. It is shown

in Section 2 that for this code we have

$$\mu\{\omega: D_F(\omega) = h\} = 1, \tag{1.3}$$

where h is the relative entropy of $\{x_n\}$ under μ. (The relative entropy is the entropy divided by lg s; see [8].) It is shown that in a special case F reduces to Fano coding. In Section 3 it is shown that if ϕ is any code then

$$\mu\{\omega: D_\phi^-(\omega) < h\} = 0. \tag{1.4}$$

Thus F achieves maximum efficiency. The methods used to establish (1.4) are those of Hausdorff dimension theory [1, 2]. In Section 4 we investigate the extent to which it is possible to replace $D_n(\omega)$ by $C_n(\omega)$ in these results. In particular, it is shown that (1.3) still holds if $D_F(\omega)$ is replaced by $C_F(\omega)$.

In [9] Kinney has exhibited a relation between Hausdorff dimension and the capacity of a noiseless channel in which the letters are of different durations. In this paper the letters are assumed all to have the same duration, so that the channel has capacity lg s.

For treatments of the noiseless coding theorem from other points of view see Feinstein [4] and Khinchine [8].

2. The Code F. As in Section 1, let $F(\alpha) = \mu(0, \alpha]$. Then, as a code, F has the desirable property that the set of ω such that $F(\omega) = F(\omega')$ for some $\omega' \neq \omega$, has μ-measure 0. In other words the original sequence can be recovered from the encoded sequence, with probability one.

THEOREM 2.1. *If μ is atomless, stationary and ergodic, then*

$$\mu\{\omega: D_F(\omega) = h\} = 1, \tag{2.1}$$

where h is the relative entropy of $\{x_n\}$ under μ.

PROOF. Since $F(\alpha) = \mu(0, \alpha]$ and μ is atomless, $\lambda(F(u)) = \mu(u)$ for any interval u. (This is just a paraphrase of the assertion that if X is a random variable with continuous distribution function $H(x)$, then $H(X)$ is a random variable which is uniformly distributed on the unit interval.) Therefore

$$D_n(\omega) = -n^{-1} \lg_s \mu(u_n(\omega)).$$

And now (2.1) follows immediately from Breiman's version of the Shannon-McMillan theorem [3].

Note that the coded process, defined by $F(\omega) = \sum_{n=1}^{\infty} y_n(\omega)/s^n$, is independent and satisfies $\mu\{\omega: y_n(\omega) = i\} \equiv 1/s$. From this it follows that F does not commute with the shift $T\omega = [s\omega]$, that is, $F(T\omega)$ and $TF(\omega)$ are in general distinct. For otherwise the processes $\{x_n\}$ and $\{y_n\}$ would be conjugate (see [6]), which they are not (unless $F(\omega) \equiv \omega$).

Note also that since $D_n(\omega)$ converges to h in $L_1(\mu)$, we have $\int_\Omega D_n(\omega)\mu(d\omega) \to h$, so that the *average* compression is also h in the limit.

3. The General Code. We now show that the code F of the preceding section is optimal in the sense that no code ϕ has a compression ratio smaller than h. Specifically, we have the following result.

THEOREM 3.1. *If μ is atomless, stationary and ergodic, and if ϕ is any code, then*

$$\mu\{\omega: D_\phi^-(\omega) < h\} = 0, \tag{3.1}$$

where h is the relative entropy of $\{x_n\}$ under μ.

PROOF. Let ν be the probability measure on $\mathfrak{B}$ with ϕ as its distribution function. Since ϕ is continuous, ν has no atoms and for any interval u, $\nu(u) = \lambda(\phi(u))$. Therefore

$$D_n(\omega) = -\frac{1}{n}\lg_s \nu(u_n(\omega)) = \frac{\lg_s \nu(u_n(\omega))}{\lg_s \mu(u_n(\omega))}\left\{-\frac{1}{n}\lg_s \mu(u_n(\omega))\right\}.$$

Since the second factor on the right goes to h almost everywhere, we have

$$D_\phi^-(\omega) = h \liminf_{n\to\infty} \frac{\lg \nu(u_n(\omega))}{\lg \mu(u_n(\omega))}$$

except on a set of μ-measure 0. Therefore, in order to prove (3.1), it suffices to show that if $\theta < h$ then

$$\mu\left\{\omega: \liminf_{n\to\infty} \frac{\lg \nu(u_n(\omega))}{\lg \mu(u_n(\omega))} \leqq \frac{\theta}{h}\right\} = 0. \tag{3.2}$$

We prove (3.2) by using results from the theory of Hausdorff dimension. For any set $M \subset \Omega$ and probability measure μ on $\mathfrak{B}$ we define the dimension $\dim_\mu M$ of M relative to μ in the following way. Consider a sum $\sum_i \mu(v_i)^\alpha$, where $\alpha > 0$ and $\{v_i\}$ is a collection of s-adic intervals covering M (that is, $M \subset \bigcup_i v_i$), with $\mu(v_i) < \rho$ for each i. The infimum of such sums we denote by $L_\mu(M, \alpha, \rho)$. As ρ decreases to 0, $L_\mu(M, \alpha, \rho)$ increases to a limit $L_\mu(M, \alpha)$, and $\dim_\mu M$ is defined by

$$\dim_\mu M = \sup\{\alpha: L_\mu(M, \alpha) = \infty\} = \inf\{\alpha: L_\mu(M, \alpha) = 0\}.$$

See [1, 2] for the details. If μ is Lebesgue measure, then $\dim_\mu M$ is the classical Hausdorff dimension of M (see [1, Section 3]).

The result relevant to coding theory is the following one (Theorem 2.1 of [2]).

If μ and ν are probability measures on $\mathfrak{B}$, and if μ is atomless, then

$$\dim_\mu\left\{\omega: \liminf_{n\to\infty} \frac{\lg \nu(u_n(\omega))}{\lg \mu(u_n(\omega))} \leqq \delta\right\} \leqq \delta. \tag{3.3}$$

Applying this result with $\delta = \theta/h$ we have

$$\dim_\mu\left\{\omega: \liminf_{n\to\infty} \frac{\lg \nu(u_n(\omega))}{\lg \mu(u_n(\omega))} \leqq \frac{\theta}{h}\right\} < 1$$

if $\theta < h$. But any set of positive μ-measure has μ-dimension 1. This proves (3.2) and the theorem.

It is possible to prove (3.2) without explicitly introducing the notion of Hausdorff dimension. This is somewhat unsatisfactory since it removes the arguments used to establish (3.3) from their natural context. In any case, the proof goes as

follows. Take $\theta/h = 1 - \epsilon$ and let A be the set of ω for which $\nu(u_n(\omega)) \geqq \mu(u_n(\omega))^{1-\epsilon}$ for infinitely many n. Since the set in brackets in (3.2) is contained in A, it suffices to show that $\mu(A) = 0$. Let ρ be an arbitrarily small positive number and let $\mathfrak{U}$ be the set of those s-adic intervals $u_n(\omega)$, with $\omega \in A$, for which $\mu(u_n(\omega)) < \rho$ and $\nu(u_n(\omega)) \geqq \mu(u_n(\omega))^{1-\epsilon}$. From the definition of A and the fact that μ is atomless it follows that the elements of $\mathfrak{U}$ cover A. Let $\mathfrak{V}$ consist of those elements of $\mathfrak{U}$ which are not subsets of other elements of $\mathfrak{U}$. Then the collection $\mathfrak{V}$ of s-adic intervals cover A and is disjoint. Moreover $\mu(v) < \rho$ and $\nu(v) \geqq \mu(v)^{1-\epsilon}$ for any $v \in \mathfrak{V}$. Therefore

$$1 \geqq \sum_{v \in \mathfrak{V}} \nu(v) \geqq \sum_{v \in \mathfrak{V}} \mu(v)^{1-\epsilon} \geqq \rho^{-\epsilon} \sum_{v \in \mathfrak{V}} \mu(v) \geqq \rho^{-\epsilon}\mu(A).$$

Thus $\mu(A) \leqq \rho^{\epsilon}$ for any $\rho > 0$, and it follows that $\mu(A) = 0$. The point is that if ω lies in A then the function ϕ is increasing very rapidly at ω, and if $\mu(A)$ were positive, the function ϕ could not remain bounded.

In Section 1 we made the assumption that $\phi(0) = 0$ and $\phi(1) = 1$. If we only assume $0 \leqq \phi(0) \leqq \phi(1) \leqq 1$, we can define ν by $\nu(0, \alpha] = \phi(\alpha) - \phi(0)$. Then ν will be a finite measure, though not a probability measure, and the above argument still goes through.

Since $D_{\phi}^{-}(\omega) \geqq h$ except on a set of μ-measure 0, an application of Fatou's lemma yields

$$\liminf_{n \to \infty} \int_{\Omega} D_n(\omega)\mu(d\omega) \geqq h. \tag{3.4}$$

Thus h is the minimal *average* compression as well as the minimum in the sense of (3.1). With somewhat different definitions and assumptions, Khinchine has proved (3.4) with the limit inferior replaced by a limit superior (see pp. 23 ff. of [8]).

As an example suppose that $s = 4$ and that under μ the process $\{x_n\}$ is independent with $\mu\{\omega: x_n(\omega) = i\} = p_i$, where $p_0 = \frac{1}{2}$, $p_1 = \frac{1}{4}$, and $p_2 = p_3 = \frac{1}{8}$. Fano coding (see [10]) proceeds in the following manner. Each symbol x_n in the x-sequence is replaced by a set of 0's and 1's according to the following rule.

$$\begin{aligned} 0 &\to 0 \\ 1 &\to 10 \\ 2 &\to 110 \\ 3 &\to 111. \end{aligned}$$

Thus $(x_1, x_2, \cdots)$ is replaced by a sequence of binary digits. These digits are then grouped in two's and put in base four again by the rule

$$\begin{aligned} 00 &\to 0 \\ 01 &\to 1 \\ 10 &\to 2 \\ 11 &\to 3. \end{aligned}$$

For example, $(0, 1, 3, 2, 0, \cdots)$ becomes $(1, 1, 3, 3, 0, \cdots)$. If, for each ω, the sequence $(x_1(\omega), x_2(\omega), \cdots)$ is transformed in this fashion into a sequence $(y_1, y_2, \cdots)$, and if we define $\phi(\omega) = \sum_{n=1}^{\infty} y_n/4^n$, then it is not difficult to show that ϕ is just the distribution function corresponding to the measure μ defined above. Therefore, by Theorems 2.1 and 3.1, the code ϕ is optimal and achieves a compression of $h = \frac{7}{8}$, as is well known.

In the preceding example we showed that a code was optimal for a process by observing that, viewed as a function on [0, 1], it is the distribution function corresponding to the process. As a second example, we construct an optimal code by starting from the distribution function. Suppose that $s = 3$ and that $\{x_n\}$ is independent with $\mu\{\omega: x_n(\omega) = i\} = p_i$, $p_0 = p_2 = \frac{1}{2}$, $p_1 = 0$. In this case the corresponding distribution function is just the Cantor function. Therefore the optimal coding rule is to replace each 2 in the x-sequence by a 1 and to convert the resulting sequence of 0's and 1's, viewed as a binary expansion, to base 3. The resulting compression ratio, lg 2/lg 3, is just that achieved by converting from base 2 to base 3.

4. Replacement of D_n by C_n. In Section 2 and Section 3 we used $D_n(\omega)$, as defined by (1.2), instead of $C_n(\omega)$, as defined by (1.1), to simplify the mathematics. Now $C_n(\omega)$ and $D_n(\omega)$ have the same asymptotic properties for any ω for which

$$\lim_{n\to\infty} \frac{\lg_s \lambda\langle\phi(u_n(\omega))\rangle}{\lg_s \lambda(\phi(u_n(\omega)))} = 1. \tag{4.1}$$

Fix ω and let $y = \phi(\omega)$, $(y - \epsilon_n, y + \delta_n] = \phi(u_n(\omega))$, and let v_n be the smallest s-adic interval containing $(y - \epsilon_n, y + \delta_n]$. We will first determine conditions on y, in terms of its nonterminating base s expansion $y = \sum_{n=1}^{\infty} y_n/s^n$, which ensure that

$$\lim_{n\to\infty} \frac{\lg_s \lambda(v_n)}{\lg_s (\epsilon_n + \delta_n)} = 1. \tag{4.2}$$

For each n, let $N_n = N_n(y)$ be the length of the run of either 0's or $(s - 1)$'s following y_n in the expansion of y. That is, determine N_n by the requirement that either

$$y_{n+1} = y_{n+2} = \cdots = y_{n+N_n} = 0 \neq y_{n+N_n+1}$$

or else

$$y_{n+1} = y_{n+2} = \cdots = y_{n+N_n} = s - 1 \neq y_{n+N_n+1}.$$

If y_{n+1} is neither 0 nor $s - 1$ then $N_n = 0$.

THEOREM 4.1. *If* $\epsilon_n \downarrow 0$, $\delta_n \downarrow 0$, $\epsilon_n + \delta_n > 0$, *and if* $\lim_{n\to\infty} N_n(y)/n = 0$ *then* (4.2) *holds.*

PROOF. Since $\lg_s \lambda(v_n)/\lg_s (\epsilon_n + \delta_n) \leqq 1$, it suffices to show that

$$\liminf_{n\to\infty} \frac{\lg_s \lambda(v_n)}{\lg_s (\epsilon_n + \delta_n)} \geqq 1.$$

The s-adic interval v_n can be determined in the following way. Let $\nu(n)$ be the largest integer ν such that

$$(y - \epsilon_n, y + \delta_n] \subset \left(\sum_{i=1}^{\nu} \frac{y_i}{s^i}, \sum_{i=1}^{\nu} \frac{y_i}{s^i} + \frac{1}{s^\nu}\right]. \tag{4.3}$$

That $\nu(n)$ is finite follows from the assumption that $\epsilon_n + \delta_n > 0$. Now v_n is just the right-hand member of (4.3) with $\nu = \nu(n)$. Therefore $\lg_s \lambda(v_n) = -\nu(n)$ and we must prove that

$$\liminf_{n\to\infty} \frac{\nu(n)}{-\lg_s (\epsilon_n + \delta_n)} \geqq 1.$$

From the fact that $\nu(n)$ is maximal it follows that either $y - \epsilon_n < \sum_{i=1}^{\nu(n)+1} y_i/s^i$ or else $y + \delta_n > \sum_{i=1}^{\nu(n)+1} y_i/s^i + 1/s^{\nu(n)+1}$. Hence, writing $\bar{y}_i = s - 1 - y_i$, one or the other of the relations

$$\epsilon_n > \sum_{i=\nu(n)+2}^{\infty} y_i/s^i, \qquad \delta_n > \sum_{i=\nu(n)+2}^{\infty} \bar{y}_i/s^i$$

holds. But the right-hand member of each of these two inequalities is not less than $s^{-[\nu(n)+2+N_{\nu(n+2)}]}$. Therefore $-\lg_s (\epsilon_n + \delta_n) \leqq \nu(n) + 2 + N_{\nu(n)+2}$, and it suffices to prove that

$$\liminf_{n\to\infty} \frac{\nu(n)}{\nu(n) + 2 + N_{\nu(n)+2}} \geqq 1.$$

Since $\epsilon_n + \delta_n$ goes to 0, $\nu(n)$ goes to infinity as n does. Hence it is enough to show that

$$\liminf_{k\to\infty} \frac{k - 2}{k + N_k} \geqq 1.$$

But this follows immediately from the assumption that $\lim_k N_k/k = 0$.

There remains the question of the size of the y-set where $\lim_n N_n(y)/n = 0$.

THEOREM 4.2. *The set of y in the unit interval for which $\lim_{n\to\infty} N_n(y)/n = 0$ has Lebesgue measure* 1.

PROOF. Since $\lambda\{y: N_n(y) \geqq n\epsilon\} = 2s^{-[n\epsilon]}$, it follows from the Borel-Cantelli lemma that $\lambda\{y: N_n(y) \geqq n\epsilon \text{ i.o.}\} = 0$. From this the theorem follows immediately. (It is possible to prove the stronger result that $N_n(y) = 0(\lg n)$ except on a set of Lebesgue measure 0. See problem 5, p. 197 of [5].)

It follows immediately from Theorems 4.1 and 4.2 that we can replace $D_F(\omega)$ by $C_F(\omega)$ in Theorem 2.1. In fact, if U is the set of y for which $\lim_n N_n(y)/n = 0$ then $\lambda(U) = 1$. But $F(\omega) \in U$ if and only if $\omega \in F^{-1}U$, and $\mu(F^{-1}U) = \lambda(U) = 1$. Therefore (4.1) holds except for ω in a set of μ-measure 0.

Similarly we can replace $D_\phi^-(\omega)$ by $C_\phi^-(\omega)$ in Theorem 3.1, provided $\mu(\phi^{-1}U) = 1$. General conditions under which this holds seem difficult to obtain.

REFERENCES

[1] PATRICK BILLINGSLEY, "Hausdorff dimension in probability theory," *Ill. J. Math.* Vol. 4 (1960), pp. 187–209.

[2] PATRICK BILLINGSLEY, "Hausdorff dimension in probability theory II," *Ill. J. Math.*, to appear.

[3] LEO BREIMAN, "The individual ergodic theorem of information theory," *Ann. Math. Stat.*, Vol. 28 (1957), pp. 809–811; "Correction note," *Ann. Math. Stat.*, Vol. 31 (1960), pp. 809–810.

[4] AMIEL FEINSTIN, *Foundations of Information Theory*, McGraw-Hill, New York, 1958.

[5] WILLIAM FELLER, *An Introduction to Probability Theory and its Applications*, 2nd. ed., John Wiley and Sons, New York, 1957.

[6] PAUL R. HOLMOS, *Entropy in Ergodic Theory*, mimeographed notes, The University of Chicago, 1959.

[7] T. E. HARRIS, "On chains of infinite order," *Pacific J. Math.*, Vol. 5 (1955), pp. 707–724.

[8] A. I. KHINCHINE, *Mathematical Foundations of Information Theory*, Dover Pub., New York, 1957.

[9] J. R. KINNEY, "Singular functions associated with Markov chains," *Proc. Amer. Math. Soc.*, Vol. 9 (1958), pp. 603–608.

[10] C. SHANNON, "A mathematical theory of communication," *Bell System Tech. J.*, Vol. 27 (1948), pp. 379–423.