The Annals of Probability
1990, Vol. 18, No. 2, 441–452

UNIVERSAL ALMOST SURE DATA COMPRESSION

By Donald S. Ornstein[1] and Paul C. Shields[2]

Stanford University and University of Toledo

An n-code is a mapping c_n from the set A^n of sequences of length n drawn from a finite set A into the set of finite length binary sequences B^*. A decoder with distortion D is a map from B^* back into A^n that sends $c_n(a_1^n)$ into a sequence that agrees with a_1^n in all but at most Dn places. We describe a sequence of codes and associated decoders of distortion D such that, for almost every sequence from an ergodic process, the number of bits per A-symbol converges almost surely to $R(D)$, the optimal compression attainable for the process. The codes are universal in that the statistics of the process need not be known in advance. Expected value results of this type were first obtained by Davisson and, independently, Fittinghof; almost sure results for the invertible case ($D = 0$) are implicitly contained in the Ziv–Lempel algorithm. Our results also apply, virtually without change in proof, to random fields.

1. Introduction. For our purposes, a source μ is a stationary, ergodic process $\{X_n\}$ taking values in a fixed finite set A, called the alphabet. The alphabet size will also be denoted by A; for cardinalities of other sets we use the common $|\cdot|$ notation. The set of n-length sequences drawn from A will be denoted by A^n and x_m^n will denote $x_m, x_{m+1}, \dots, x_n$. A code (or more precisely, a binary n-code) is a function c_n from A^n into the set of finite length binary sequences $B^* = \bigcup_{n=1}^{\infty} B^n$, where $B = \{0, 1\}$. The length function $l(x_1^n)$ is defined by $c_n(x_1^n) = b_1^{l(x_1^n)}$, and its associated compression factor is defined by $r(c_n(x_1^n)) = l(x_1^n)/n$. The expected compression factor, $R(c_n) = E_\mu(r(c_N(x_1^n)))$, is called the *rate* of the code. The code is called a *block* code if $l(x_1^n)$ is constant, a.s.; otherwise it is called a *variable-length code*.

To measure fidelity, we use the average Hamming distance

$$d_n(x_1^n, y_1^n) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, y_i),$$

where

$$d(x, y) = \begin{cases} 1, & \text{if } x = y, \\ 0, & \text{otherwise.} \end{cases}$$

A code c_n will be called *D-semifaithful* if there is a function $\psi_n\colon B^* \to A^n$, called the *decoder*, such that

$$d_n(x_1^n, \psi_n(c_n(x_1^n))) \le D, \qquad x_1^n \in A.$$

Received September 1988; revised April 1989.

[1] Partially supported by NSF Grant DMS-86-05098.

[2] Partially supported by NSF Grant DMS-87-42630.

AMS 1980 *subject classifications*. Primary 94A34; secondary 28D20.

Key words and phrases. Universal data compression, entropy, rate-distortion function.

In the invertible case ($D = 0$), such a code will be called *faithful* or *noiseless*. The (operational) rate-distortion function $R_\mu(D)$ is defined as follows: If $\tilde{a}_1^n$ is fixed, the set $\{a_1^n \colon d_n(a_1^n, \tilde{a}_1^n) \le D\}$ will be called a (D, n)-ball with center $\tilde{a}_1^n$, or simply a D-ball if n is understood. Let $\mathscr{S}$ be a subset of A^n and define $N(D, \mathscr{S})$ to be the minimum number of D-balls needed to cover $\mathscr{S}$. Define

$$R_n(D, \varepsilon) = \min_{\mathscr{S} \colon \mu(\mathscr{S}) \ge 1 - \varepsilon} \log_2 N(D, \mathscr{S}),$$

that is, the exponent in the minimum number of D-balls needed to cover a set in A^n of probability $1 - \varepsilon$. Then define

$$R_\mu(D) = \lim_{\varepsilon \to 0} \lim_{n \to \infty} R_n(D, \varepsilon).$$

As noted in our final remark, $R_\mu(D) = \lim_{n \to \infty} R_n^*(D)$, where $R_n^*(D)$ is the greatest lower bound of the rates $R(c_n)$ over the class of *all* D-semifaithful n-codes c_n.

The landmark papers by Davisson [1], Lynch [6] and Fittinghof [4], first showed the existence of asymptotically optimal noiseless codes in the expected value sense for some classes of sources. These results were extended by many others, until it was finally established that for the class of ergodic sources with a given finite alphabet A, there is a sequence $\{c_n\}$ of D-semifaithful codes such that for each ergodic process μ the compression factor $r(c_n(x_1^n))$ converges in L^1-norm to $R_\mu(D)$ [5]. The Ziv–Lempel algorithm [11] provides a sequence $\{c_n\}$ of invertible codes such that for any ergodic source μ the sample compression ratio $r(c_n(x_1^n))$ converges almost surely to $R_\mu(0) = H$, the entropy of the process. Our principal result is an extension of the Ziv–Lempel result to obtain almost sure convergence for the semifaithful case.

THEOREM 1. *For any $D \ge 0$ there is a sequence $\{c_n\}$ of D-semifaithful codes such that for any ergodic source μ, the sample compression factor $r(c_n(x_1^n))$ converges almost surely to $R_\mu(D)$.*

If the source is known, then the definition of the rate-distortion function implies the existence of an asymptotically optimal sequence of codes. Theorem 1 asserts that a sequence of codes exists that is universal, that is, we can design a code in advance such that for any given ergodic source the code almost surely performs as well asymptotically as a code designed specifically for the source. Our proof is based on the covering ideas contained in a recent paper by Ornstein and Weiss [8]. In order to make the proof more easily understood by the reader, we first give the proof in the noiseless case, then extend to the semifaithful case.

A feature of our method of proof is that it extends to random fields, that is, processes of the form $\{x_{n_1, n_2, \dots, n_M}\}$. This extension to random fields is essentially a mere translation from one dimension to M dimensions; e.g., we just replace a single integer n by an integer-valued M-vector $n = (n_1, n_2, \dots, n_M)$ and define a block to be an M-dimensional rectangle. We note at the end of our paper the place in Section 3 where somewhat more work is needed in the random field case.

2. The invertible case. Our code will use the partition of x_1^n into the contiguous nonoverlapping blocks

$$\{x_{ik+1}^{(i+1)k}: i = 1, 2, \dots, q\}, \qquad n = kq + r, 0 \le r < k,$$

and the empirical k-block distribution $\hat{p}_k$ defined by

$$\hat{p}_k(a_1^k) = \frac{|\{i: x_{ik+1}^{(i+1)k} = a_1^k, i < (n-r)/k\}|}{(n-r)/k}. \tag{1}$$

The remainder x_{n-r+1}^n can always be encoded separately without affecting asymptotic performance since we will assume that k grows like $\frac{1}{2}\log_A n$. It is sufficient, therefore, to construct the codes for the case $n = kA^{2k}$, a relation assumed to hold in our subsequent code constructions.

An informal description of our code is the following: First, the sequence x_1^n is partitioned into contiguous nonoverlapping k-blocks. Second, a list of these k-blocks in order of decreasing frequency is transmitted, which we call the code book. This initial listing of the code book is short relative to n, since k is asymptotically $\frac{1}{2}\log_A n$. Successive k-blocks in x_1^n are then encoded by giving the index of the block in the code book. High frequency blocks appear near the front of the code book and therefore have short indices, which will guarantee optimal asymptotic performance. A construction similar in spirit to ours was used to obtain asymptotically optimal expected coding rates in [10]; our proof techniques are quite different, however, and yield almost sure convergence.

To describe our code more precisely, let us first define the concatenation operation $*$ as

$$a * b = a_1, a_2, \dots, a_s, b_1, b_2, \dots, b_t, \qquad a = a_1^s,\ b = b_1^t.$$

(Commas are used here for visual clarity.) Next we choose a fixed invertible function $f\colon A \to B^s$, where $2^{s-1} < A \le 2^s$, to encode the alphabet symbols, and then extend to sequences by concatenation to obtain $f_n\colon A^n \to B^{sn}$, defined by

$$f_n(a_1^n) = f(a_1) * f(a_2) * \cdots * f(a_n).$$

Fix a sequence x_1^n. The encoded sequence $c_n(x_1^n) = b_1^{l(x_1^n)}$ will be the concatenation of two binary sequences b_1^m and $b_{m+1}^{l(x_1^n)}$, that is,

$$c_n(x_1^n) = b_1^{l(x_1^n)} = b_1 b_2 \cdots b_m b_{m+1} \cdots b_{l(x_1^n)} = b_1^m * b_{m+1}^{l(x_1^n)}.$$

The first part b_1^m, the encoding of the code book, is defined as follows: Arrange the possible sequences a_1^k in order of decreasing values of their empirical probabilities $\hat{p}_k(a_1^k)$, defined by (1), and call this list the code book $\mathscr{L}$. Define b_1^m to be the $*$-concatenation of the binary blocks $f_k(a_1^k)$ in order of the appearance of a_1^k in the code book $\mathscr{L}$. Note that the length of this first part m is asymptotically upper bounded by $k(1 + \log_2 A)A^k$, which is $o(n)$, so that the listing of the code book makes a negligible contribution to code length.

Note also that the decoder does not need to know the actual frequencies of each k-block, but only an ordering in terms of frequencies.

The second part of the code $b_{m+1}^{l(x_1^n)}$ contains the addresses in the code book $\mathscr{L}$ of the successive k-blocks in x_1^n. To make this precise, let us define

$$\mathscr{A}(x_{ik+1}^{(i+1)k}) = j, \quad \text{if } x_{ik+1}^{(i+1)k} \text{ is the } j\text{th sequence in the list } \mathscr{L}.$$

To obtain compression these addresses must be expressed as binary sequences of variable lengths so that high frequency blocks, which appear at the beginning of $\mathscr{L}$, will have short binary addresses. To guarantee unique decodability of these binary addresses we use a technique of Elias [2], which was suggested to us by Imre Csiszár. A short binary prefix is added to the standard binary representation of each integer so that no binary address is a prefix of any other binary address. This is done as follows. Let BIN(n) be the usual binary representation of the integer n and let LEN(n) be the length of this binary representation of n, so that $\mathrm{BIN}(n) = b_1^{\mathrm{LEN}(n)}$. Also, let $Z(b_1^m)$ be a sequence of 0's of length m equal to the length of the binary sequence b_1^m. Then define the function $G(n)$ by concatenation as

$$G(n) = Z(\mathrm{BIN}(\mathrm{LEN}(n)))*\mathrm{BIN}(\mathrm{LEN}(n))*\mathrm{BIN}(n).$$

Note that the leading bit in the standard representation of an integer is always a 1, so that decoding is easy. The initial block of 0's must be $Z(\mathrm{BIN}(\mathrm{LEN}(n)))$. This tells how many further bits to read to determine the second block BIN(LEN(n)). This second block tells us how many further bits to read to determine BIN(n) and hence n. Thus any concatenation $G(n_1)*G(n_2)*\cdots*G(n_k)$ can be uniquely decoded to produce the sequence of addresses $n_1, n_2, \dots, n_k$. We now define $b_{m+1}^{l(x_1^n)}$ to be the concatenation of the $G(\mathscr{A}(x_{ik+l}^{(i+1)k}))$ for $0 \leq i < n/k$.

As noted earlier, the amount of space m needed to transmit the code book $\mathscr{L}$ is $o(n)$. Likewise, it is easy to see that the total space taken up by the address prefixes is also $o(n)$. Let $\mathscr{S}_n = \Sigma_i \mathrm{LEN}(\mathscr{A}(x_{ik+1}^{(i+1)k}))$. To complete the proof of Theorem 1 in the noiseless case it is enough to show that

$$\lim_{n\to\infty} \frac{\mathscr{S}_n}{n} = H, \quad \text{a.s.},$$

where H is the entropy of the process. We make use of a recent covering result of Ornstein and Weiss [8, Section 2, Theorem 2], stated as follows.

LEMMA 1. *Let $\{X_n\}$ be an ergodic process with finite alphabet A and entropy H. Let ε be a positive number. There is a K_0 such that if $k \geq K_0$ there is a collection $\mathscr{T}_k = \mathscr{T}_k(\varepsilon) \subseteq A_k$ of cardinality at most $2^{(H+\varepsilon)k}$ such that for almost every realization $x = \{x_j\}$ there is an integer $K \geq K_0$ such that if $k \geq K$ and $n \geq 2^{Hk}$, then the following are true.*

(i) *In the sequence x_1^n, the k-blocks $x_{ik+1}^{(i+1)k}$ belong to $\mathscr{T}_k$ for all but at most $(1-\varepsilon)\%$ of the indices i.*

(ii) *If $\mathscr{B}_k$ is a collection of sequences of length k such that in x_1^n the blocks $x_{ik+1}^{(i+1)k}$ belong to $\mathscr{B}_k$ for more than $\varepsilon\%$ of the indices i, then the cardinality of $\mathscr{B}_k$ is at least $2^{(H-\varepsilon)k}$.*

A proof of an extension of this lemma to the rate distortion case will be given in the next section. The lemma is applied in the following way: Let us choose $k \geq K$ and put $n = kA^{2k}$, noting that $n \geq 2^{Hk}$, since $H \leq \log_2 A$. Let $\mathscr{T}_k = \mathscr{T}_k(\varepsilon)$ be the collection of typical sequences given by Lemma 1 and let $\mathscr{T}$ be the first $2^{(H+\varepsilon)k}$ members of the code list $\mathscr{L}$, so that $|\mathscr{T}| \geq |\mathscr{T}_k|$ and $\hat{p}_k(\mathscr{T}_k) \geq 1 - \varepsilon$. Since $\mathscr{L}$ lists the blocks in order of decreasing frequency, we have $\hat{p}_k(\mathscr{T}) \geq \hat{p}_k(\mathscr{T}_k)$. Thus all but $(1 - \varepsilon)\%$ of the binary addresses $\mathrm{BIN}(\mathscr{A}(x_{ik+l}^{(i+1)k}))$ refer to members of $\mathscr{T}$ and thus have lengths bounded above by $(H + \varepsilon)k$. For the lengths of the binary addresses of those k-blocks that are not in $\mathscr{T}$, which are at most $\varepsilon\%$ of the addresses, we can use the crude upper bound $1 + k \log_2 A$. Thus we have

$$\mathscr{L}_n \leq (1 - \varepsilon)[n/k](H + \varepsilon)k + \varepsilon[n/k](1 + k \log_2 A)$$

and hence

$$\limsup_{n \to \infty} \mathscr{L}_n/n \leq H, \quad \text{a.s.}$$

To show that this limit superior is really a limit and equal a.s. to H, we make use of the second part of Lemma 1. Let $\mathscr{B}$ be the first $2^{(H-\varepsilon)k}$ sequences a_1^k in the list $\mathscr{L}$. The lemma guarantees that there can be at most $[n/k]\varepsilon$ indices i for which $x_{ik+1}^{(i+1)k}$ belongs to $\mathscr{B}$. These are the only k-blocks that can have addresses shorter than $(H - \varepsilon)k$; hence

$$\liminf_{n \to \infty} \mathscr{L}_n/n \geq H(1 - \varepsilon)(1 - \varepsilon), \quad \text{a.s.}$$

The proof of Theorem 1 in the noiseless case is now complete.

3. The semifaithful case. We now extend our previous construction to the semifaithful case. The distortion level D will be fixed throughout this section and we set $R = R_\mu(D)$. For convenience we modify our definition of D-ball to allow subsets, that is, a nonempty subset of $\{a_1^k\colon d_k(a_1^k, \tilde{a}_1^k) \leq D\}$ will be called a (D, k)-ball with center $\tilde{a}_1^k$, or simply a D-ball if k is understood. Thus, in particular, a D-ball need not include its center.

As before, we assume that $n = kA^{2k}$ and count frequencies of nonoverlapping k-blocks in x_1^n, letting $\hat{p}_k$ denote the empirical distribution of k-blocks defined by (1). Thus to say that a collection $\mathscr{C}$ of k-blocks covers $\alpha\%$ of x_1^n is to say that $\hat{p}_k(\mathscr{C}) = \alpha$. Let $\mathscr{U}_k = \mathscr{U}_k(x_1^n)$ be the set of all k-blocks that appear in x_1^n, that is,

$$\mathscr{U}_k = \{a_1^k\colon \hat{p}_k(a_1^k) > 0\}. \tag{2}$$

We construct our code book as follows: By induction, pick a sequence $\{V_i\}$ of D-balls such that V_i is the D-ball contained in $\mathscr{U}_k - \bigcup_{j<i} V_j$ of largest $\hat{p}_k$ probability. Let $a(i)_1^k$ be the center of V_i and let $\mathscr{L}(D)$ be the list of the $a(i)_1^k$ in order of increasing i. The first part of our code, b_1^m, is defined to be the concatenation of the binary blocks $f_k(a(i)_1^k)$ in order of increasing i, where, as before, $f_k\colon A^k \to B^{ks}$ is our fixed binary encoder. The second part of the code,

$b_{m+1}^{l(x_1^n)}$, is the concatenation of the $G(\mathscr{A}_D(x_{ik+1}^{(i+1)k}))$, where $\mathscr{A}_D(x_{ik+1}^{(i+1)k})$ is defined to be j if $x_{ik+1}^{(i+1)k} \in V_j$.

As in the noiseless case, the listing of the code book and the sum of all the prefix lengths are each $o(n)$, so it is enough to show that $\mathscr{S}_n/n \to R$, a.s., where $\mathscr{S}_n$ is the sum of the lengths $\mathrm{LEN}(\mathscr{A}_D(x_{ik+l}^{(i+1)k}))$. To accomplish this we again make use of the Ornstein–Weiss result (our Lemma 1), modified here so that it covers the semifaithful case. We first define the D-neighborhood of a set $\mathscr{C}$ of k-sequences as

$$\mathscr{N}_D(\mathscr{C}) = \{a_1^k \colon d(a_1^k, u_1^k) \le D, \text{ for some } u_1^k \in \mathscr{C}\}.$$

Lemma 2. *Let $\{X_n\}$ be an ergodic process with finite alphabet A, measure μ and entropy H, and set $R = R_\mu(D)$, where D is fixed. Fix a positive number ε. There is a K_0 such that if $k \ge K_0$ there is a collection $\mathscr{C}_k \subset A^k$ of centers of cardinality at most $2^{(R+\varepsilon)k}$ such that for almost every realization $x = \{x_j\}$ there is an integer $K = K(x) \ge K_0$ such that if $k \ge K$ and $n \ge 2^{Hk}$ then the following are true.*

(i) *In the sequence x_1^n, the k-block $x_{ik+1}^{(i+1)k}$ belongs to $\mathscr{N}_D(\mathscr{C}_k)$ for all but at most $(1-\varepsilon)\%$ of the indices i.*

(ii) *If $\mathscr{B}_k$ is a collection of sequences of length k such that in x_1^n the blocks $x_{ik+1}^{(i+1)k}$ belong to $\mathscr{N}_D(\mathscr{B}_k)$ for more than $\varepsilon\%$ of the indices i, then the cardinality of $\mathscr{B}_k$ is at least $2^{(R-\varepsilon)k}$.*

We first show how Lemma 2 leads to the conclusion that $\mathscr{S}_n/n \le R$, a.s. As in (1) and (2), let $\hat{p}_k$ be the empirical distribution and $\mathscr{U}_k$ the universe of k-blocks determined by x_1^n. Also let $\{V_i\}$ be the partition of $\mathscr{U}_k$ into disjoint D-balls used to define the code book $\mathscr{S}(D)$. Define M to be the first integer such that $\hat{p}_k(V_M) < 2^{-k(R+\varepsilon)}$ and define

$$\mathscr{E} = \bigcup_{i<M} V_i, \qquad \tilde{\mathscr{E}} = \bigcup_{i\ge M} V_i.$$

Our goal is to show that $\mathscr{E}$ covers most of x_1^n, that is, that $\hat{p}_k(\tilde{\mathscr{E}})$ is small, for since $M \le 2^{k(R+\varepsilon)}$, this will show that $\mathscr{S}_n/n$ cannot be much larger than $R + \varepsilon$.

We can assume that k is so large that $2^{-k\varepsilon/2} < \varepsilon/2$ and so that for $n = kA^{2k}$ Lemma 2 provides us with a collection $\mathscr{T}_k \subset \mathscr{U}_k$ such that $\hat{p}_k(\mathscr{T}_k) \ge 1 - \varepsilon/2$ and $\mathscr{T}_k$ can be partitioned into a collection $\{U_i\}$ of no more than $2^{k(R+\varepsilon/2)}$ D-balls. From the definition of M we have that

$$\hat{p}_k(U_i \cap \tilde{\mathscr{E}}) \le 2^{-k(R+\varepsilon)},$$

so that

$$\begin{aligned}\hat{p}_k(\tilde{\mathscr{E}}) &= \hat{p}_k((\mathscr{U}_k - \mathscr{T}_k) \cap \tilde{\mathscr{E}}) + \hat{p}_k(\mathscr{T}_k \cap \tilde{\mathscr{E}})\\ &\le \varepsilon/2 + 2^{k(R+\varepsilon/2)} \cdot 2^{-k(R+\varepsilon)} < \varepsilon.\end{aligned}$$

This establishes that

$$\mathscr{S}_n \leq (1-\varepsilon)[n/k](R+\varepsilon)k + \varepsilon[n/k](1 + k\log_2 A),$$

so that $\limsup_{n\to\infty}\mathscr{S}_n/n \leq R$, a.s. An argument using the second part of Lemma 2, similar to that used in Section 2, shows that this limit superior is really a limit and equal to $R = R_\mu(D)$.

Proof of Lemma 2. We sketch the proof of the first part of the lemma, indicating where changes need to be made in the Ornstein–Weiss argument. Let δ and $\eta < \delta/2$ be positive numbers to be specified later. First we use the Shannon–McMillan theorem [3] for the rate-distortion function to obtain an integer M and a subset $\hat{\mathscr{C}} = \hat{\mathscr{C}}_M$ of A^M with the two properties:

(a) The cardinality of $\hat{\mathscr{C}}$ is at most $2^{(R+\delta/2)M}$.
(b) If μ_M is the measure on A^M defined by μ, then $\mu_M(\mathscr{N}_D(\hat{\mathscr{C}})) \geq 1 - \eta^2$.

For each positive integer $k \geq K_0$, where K_0 will be prescribed later, let $\mathscr{C}_k = \mathscr{C}_{k,M}$ be the set of all sequences a_1^k such that for all but $(1 - 3\delta)\%$ of the indices i, $0 \leq i \leq k - M$, the M-block a_{i+1}^{i+M} belongs to $\hat{\mathscr{C}}_M$. As in the Ornstein–Weiss proof, a count of the possibilities shows that for suitable choice of K_0 and δ, the cardinality of $\mathscr{C}_k$ is at most $2^{(R+\delta)k}$, for all $k \geq K_0$.

From the ergodic theorem we know that for almost all infinite sequences $x = \{x_i\}$ there is an integer $K = K(x) \geq K_0$ such that if $k \geq K$ and $n \geq k$, then all but at most $(1 - 2\eta^2)\%$ of the M-blocks $x_{iM+1}^{i(M+1)}$ in x_1^n belong to $\mathscr{N}_D(\hat{\mathscr{C}}_M)$. If k is in the range $n \geq k \geq K_0$, then, by the Markov inequality, all but at most $(1 - 2\eta)\%$ of the k-blocks $x_{ik+1}^{(i+1)k}$ must belong to $\mathscr{N}_D(\mathscr{C}_k)$, since we assumed that $\eta < \delta/2$.

This implies the first conclusion of Lemma 2.

We now turn to the proof of the second part of Lemma 2. Note that the relation between n and k is now allowed to be arbitrary, subject only to the condition that $n \geq 2^{Hk}$. As before we let $\hat{p}_k$ be the distribution of k-blocks and $\mathscr{U}_k$ the universe of k-blocks determined by x_1^n, as defined by (1) and (2).

Fix a positive number ε. Let us call a sequence x_1^n ε-bad if the second part of the lemma fails for it, that is, there is a collection $\mathscr{B}_k \subset A^k$ of cardinality less than $2^{(R-\varepsilon)k}$ such that the blocks $x_{ik+1}^{(i+1)k}$ belong to $\mathscr{N}_D(\mathscr{B}_k)$ for more than $\varepsilon\%$ of the indices i. Let $\mathscr{D}(n,\varepsilon)$ be the set of all ε-bad sequences of length n. The proof of Lemma 2 will be completed if we can show that for almost all $x = \{x_j\}$, x_1^n belongs to $\mathscr{D}(n,\varepsilon)$ for only finitely many n. To obtain better control of the bad sequences we first use Lemma 1 to choose a collection $\mathscr{T}_k \subset A^k$ of cardinality no more than $2^{k(H+\delta)}$, such that for almost every $x = \{x_j\}$ we have $\hat{p}_k(\mathscr{T}_k) > 1 - \delta$, so long as k is large enough and $n \geq 2^{Hk}$. The number δ will be specified later. We think of $\mathscr{T}_k$ as the set of typical k-blocks. Now define

$$\mathscr{D}_1(n,\varepsilon) = \mathscr{D}(n,\varepsilon) \cap \{x_1^n : \hat{p}_k(\mathscr{T}_k) > 1 - \delta\}.$$

From Lemma 1 it is enough to show that for almost all $x = \{x_j\}$, x_1^n belongs to $\mathscr{D}_1(n,\varepsilon)$ for only finitely many n. Our desired result will be a consequence of the following two lemmas.

LEMMA 3. *Given $\varepsilon > 0$ there is an $\hat{\varepsilon} > 0$ such that for all sufficiently large n, $\mathscr{Q}_1(n, \varepsilon)$ can be covered by fewer than $2^{n(R-\hat{\varepsilon})}(D, n)$-balls.*

LEMMA 4. *The set A^n can be partitioned into two sets $\mathscr{F}_n$ and $\mathscr{G}_n$ such that the following hold.*

(i) *For almost every $x = \{x_j\}$, the finite sequence x_1^n belongs to $\mathscr{F}_n$ for only finitely many n.*
(ii) *No (D, n)-ball contained in $\mathscr{G}_n$ has probability larger than $2^{-n(R-\varepsilon)}$.*

The second part of Lemma 2 follows easily from these two lemmas. Lemma 3 and part (ii) of Lemma 4, with ε replaced by $\hat{\varepsilon}/2$, imply that $\mathscr{Q}_1(n, \varepsilon) \cap \mathscr{G}_n$ has probability less than $2^{-n\hat{\varepsilon}/2}$ and this together with part (i) of Lemma 4 shows that indeed $x_1^n \in \mathscr{Q}(n, \varepsilon)$ only finitely often, almost surely. Thus it is enough to prove Lemma 3 and Lemma 4.

PROOF OF LEMMA 3. Let us write $n = km + r$, where $0 \leq r < k$, and let I be the set of integers in the interval $[0, m)$. A collection $\mathscr{S}$ of k-sequences that can be covered by fewer than $2^{k(R-\varepsilon)}$ D-balls will be called *too thin*. If a sequence x_1^n is bad, then we can find a subset I_1 of I, of cardinality at least $m\varepsilon$ for which the set of sequences

$$\mathscr{S}(x_1^n, I_1) = \{x_{ik+1}^{(i+1)k} : i \in I_1\}$$

is too thin.

Our proof of Lemma 3 will use the following ideas, stated here in somewhat vague form.

(a) The number of too thin collections determined by the bad n-sequences is exponentially small in n.

(b) The set of sequences with the same too thin collection $\mathscr{S}$ covering a fixed set of places of cardinality at least $n\varepsilon$ can be covered by exponentially fewer than $2^{nR}(D, n)$-balls.

Let us begin by bounding the number of too thin collections $\mathscr{S}$ that are needed for all the bad n-sequences. Toward this end we need consider only the too thin subsets of the set $\mathscr{T}_k$ of typical k-blocks.

Next we use an idea of [3] to control the sizes of D-balls. A (D, k)-ball will be called *big* if its cardinality exceeds $2^{k(H-R+\varepsilon/4)}$. Choose a maximal set $\{V_i\}$ of disjoint big D-balls contained in $\mathscr{T}_k$ and let $\mathscr{B} = \bigcup_i V_i$. Note that since $|\mathscr{T}_k| \leq 2^{(H+\delta)k}$ the set $\mathscr{B}$ can be covered by fewer than $2^{k(R+\delta-\varepsilon/4)}$ D-balls. We can therefore assume that δ is chosen so small that for any too thin subset $\mathscr{S}$ of $\mathscr{T}_k$ the following holds:

(3) The set $\mathscr{S} \cup \mathscr{B}$ can be covered by fewer than $2^{k(R-\varepsilon/8)}$ D-balls.

The number of possible $\mathscr{S} \cup \mathscr{B}$ can be bounded as follows: Let $\tilde{\mathscr{B}} = \mathscr{T}_k - \mathscr{B}$ and note that the number of distinct $\mathscr{S} \cup \mathscr{B}$ is the same as the number of distinct $\mathscr{S} \cap \tilde{\mathscr{B}}$. To count the latter recall that, by definition, a D-ball contained in $\tilde{\mathscr{B}}$ must have fewer than $2^{k(H-R+\varepsilon/4)}$ members, so that since a too thin collection $\mathscr{S}$ can be covered by fewer than $2^{(R-\varepsilon)k}$ (D,k)-balls, we have

$$|\mathscr{S}\cap \tilde{\mathscr{B}}| \leq 2^{k(R-\varepsilon)}2^{k(H-R+\varepsilon/4)} \leq 2^{k(H-\varepsilon/2)}.$$

Therefore, if $n \geq 2^{kH}$, the number of possible sets $\mathscr{S}\cup \mathscr{B}$ is upper bounded by

$$(4)\quad \binom{2^{k(H+\delta)}}{2^{k(H-\varepsilon/2)}} \leq 2^{2k(\varepsilon+\delta)2^{-k(\varepsilon+\delta)}2^{k(H+\delta)}} \leq 2^{n\beta n}, \quad \text{where } \beta_n \to 0 \text{ as } n \to \infty,$$

where we used the fact that $\binom{n}{pn} \leq 2^{-2np\log p}$ if $p < \frac{1}{2}$.

Now we show how to cover the bad set $\mathscr{Q}_1(n,\varepsilon)$ by exponentially fewer than 2^{Rn} (D,n)-balls. By using the first part of Lemma 2 we can assume

(5) $\mathscr{T}_k$ can be covered by fewer than $2^{k(R+\delta)}$ (D,k)-balls.

Fix a set $I_0 \subset I$ of cardinality at most δ, fix a set $I_1 \subset I$ of cardinality at least $m\varepsilon$, and fix a too thin set $\mathscr{S}$ contained in $\mathscr{T}_k$. Consider the set $\mathscr{Q}(\mathscr{S}, I_1)$ of all bad sequences x_1^n such that $\mathscr{S}(x_1^n, I_1) \subset \mathscr{S}\cup \mathscr{B}$ and the indices i for which $x_{ik+1}^{(i+1)k} \notin \mathscr{T}_k$ all belong to I_0. We cover $\mathscr{Q}(\mathscr{S}, I_1)$ by (D,n)-balls by using the (D,k)-balls that cover $\mathscr{S}\cup \mathscr{B}$ to cover the k-blocks that start in I_1, then using the (D,k)-balls that cover $\mathscr{T}_k$ to cover the k-blocks that start in $I - I_0 - I_1$ and finally covering the remaining k-blocks and the last r terms in some arbitrary way. Thus, using (3) and (5), the number of (D,n)-balls needed to cover $\mathscr{Q}(\mathscr{S}, I_1)$ is bounded above by

$$(6)\qquad 2^{(R-\varepsilon/8)k|I_1|} \cdot 2^{(R+\delta)k(m-\delta m-|I_1|)} \cdot A^{k\delta m} \cdot A^r \leq 2^{(R-\varepsilon_1)n}$$

if δ is small enough, where ε_1 is positive and independent of n. Since the number of subsets of I, the set of places where k-blocks can start, is upper bounded by $2^m \leq 2^{n/k}$, we also have:

(7) The number of ways to choose I_1 and I_0 is upper bounded by $2^{2n/k}$.

Thus the bounds (7), (4) and (6), with δ chosen small enough and k large enough show that there is a positive number $\hat{\varepsilon}$ such that the bad set $\mathscr{Q}_1(n,\varepsilon)$ can be covered by fewer than $2^{n(R-\hat{\varepsilon})}$ (D,n)-balls. This completes the proof of Lemma 3. □

Proof of Lemma 4. We again make use of the idea suggested in [3], this time to control the measures of D-balls. Let us call a (D,n)-ball *fat* if its measure is at least $2^{-n(R-\varepsilon)}$. Note that fatness refers to measure, while bigness referred to cardinality. Let $\Phi = \{\mathscr{V}_i\}$ be a maximal disjoint collection of fat (D,n)-balls and put $\mathscr{F}_n = \bigcup_i \mathscr{V}_i$, $\mathscr{G}_n = A^n - \mathscr{F}_n$. Then property (ii) of Lemma 4 certainly holds.

Note that the fat part of the space $\mathscr{F}_n$ can be covered by fewer than $2^{n(R-\varepsilon)}$ (D,n)-balls. This is the only property of the sequence $\{\mathscr{F}_n\}$ that will be used in the remainder of the proof. To make this explicit we state this fact as the following lemma, from which Lemma 4 follows easily.

LEMMA 5. *Let $\{\mathscr{F}_n\}$ be a sequence of sets such that for each n, the set $\mathscr{F}_n$ can be covered by fewer than $2^{n(R-\varepsilon)}$ (D, n)-balls. Then for almost every sequence $x = \{x_i\}$, the block x_1^n belongs to $\mathscr{F}_n$ for at most finitely many n.*

PROOF. We shall show that if Lemma 5 is false, then for sufficiently large m we can cover a set of m-sequences of large probability with exponentially fewer than 2^{mR} (D, m)-balls. To establish this we shall use the covering argument recently employed to extend the asymptotic equipartition property to amenable groups in [7] (see also the appendix in [9] for a description of the one-dimensional form of the construction used here).

Let us suppose that Lemma 5 is false, so there is a positive number γ such that for any N there is an $M > N$ such that the set

$$W_N^M = \bigcup_{n=N}^{n=M} \{x : x_1^n \in \mathscr{F}_n\}$$

has measure greater than γ. Let δ be a positive number to be specified later and use the definition of the rate-distortion function to find a positive integer L and a collection $\mathscr{C}_L$ of L-sequences of probability greater than $1 - \delta$ that can be covered by fewer than $2^{L(R+\delta)}$ D-balls. Choose N so large that $2L/(N + 2L) < \delta$ and then choose M so that $\mu(W_N^M) > \gamma$. Put $\tilde{W} = \bigcup_{n=N}^{n=M} \mathscr{F}_n$. Note that $\tilde{W}$ is a set of finite sequences of lengths varying between N and M, while $W_N^M = \{x\colon x_1^n \in \tilde{W}$, for some n, $N \le n \le M\}$.

Let us call x_1^n *good* if we can find a collection $\mathscr{U}$ of nonoverlapping blocks x_{i+1}^{i+m} with the properties:

(a) The collection $\mathscr{U}$ is the union of two disjoint subcollections, $\mathscr{C}$ and $\mathscr{W}$, such that the blocks in $\mathscr{C}$ all have length L and belong to $\mathscr{C}_L$ while the blocks in $\mathscr{W}$ all belong to $\tilde{W}$.

(b) $\mathscr{U}$ covers all but $2\delta\%$ of x_1^n.

(c) $\mathscr{W}$ covers at least $\gamma\%$ of x_1^n.

Let Γ_n be the set of all good n-sequences. We shall prove

$$\lim_{n\to\infty} \mu(\Gamma_n) = 1. \tag{8}$$

We sketch the covering argument used to prove this in the one-dimensional case, referring the reader to the appendix in [9] for the details. Let T denote the shift on sequences $x = \{x_j\}$. For a given x define two increasing sequences $\{n_i\}$ and $\{m_i\}$ of positive integers with $n_i < m_i < n_{i+1}$ as follows: Let n_1 be the first positive integer such that $x_{n_i}^m$ belongs to $\tilde{W}$ and define m_1 to be the least such m. Having defined n_i and m_i for $i < j$, define n_j to be the least integer greater than m_{j-1} such that there is an m such that $x_{n_j}^m \in \tilde{W}$, and define m_j to be the least such m. The process stops as soon as we get within $M - N$ of the end of x_1^n. Since the set of indices i for which $x \notin T^i W_N^M$ has limiting density less than $1 - \gamma$, this proves that for all sufficiently large n, for most sequences of length n we can cover at least $\gamma\%$ of the sequence with nonoverlapping blocks that belong to W. Here "most" means in terms of probability.

By a similar argument we can eventually also cover all but $\delta\%$ of most sequences of length n by nonoverlapping L-blocks that belong to $\mathscr{C}_L$. Combining these results we see that for n large we can with high probability both cover $\gamma\%$ of x_1^n by nonoverlapping blocks from $\tilde{W}$ and cover $(1 - \delta)\%$ by nonoverlapping L-blocks from $\mathscr{C}_L$. The nonoverlapping blocks that belong to $\tilde{W}$ are assigned to $\mathscr{W}$ and the nonoverlapping L-blocks that are in $\mathscr{C}_L$ and that meet no blocks in $\mathscr{W}$ are assigned to $\mathscr{C}$. Since we assumed that $2L/(N + 2L) < \delta$ we will with high probability in this way obtain good n-blocks. This proves (8).

From the definition of fatness, the cardinality of the maximal set of fat (D, n)-balls is at most $2^{n(R-\varepsilon)}$. Thus we can apply to Γ_n an argument similar to the one used to prove Lemma 3 to show that there is an $\alpha > 0$ such that for all sufficiently large n, Γ_n can be covered by fewer than $2^{n(R-\alpha)}$ D-balls. This contradicts the assumption that $R = R(D)$ and establishes Lemma 5. Our proof of Theorem 1 is now complete. □

REMARK 1. Note that our proof of the result (8) made use of the linear ordering of the natural numbers. For random fields in higher dimensions the nesting argument used in [7] can be used to obtain the same conclusion.

REMARK 2. Lemma 5 also shows that there is no way to beat the rate-distortion function in the limit, no matter what sequence of D-semifaithful codes $\{c_n\}$ is used. To see this let $R = R_\mu(D)$ and suppose that there is a sequence of codes such that

$$\limsup_{n\to\infty} c_n(x_1^n) < R, \quad \text{almost surely}.$$

Then for some $\varepsilon > 0$ the sets

$$\mathscr{B}(n, \varepsilon) = \{x_1^n : c_n(x_1^n) \le R - \varepsilon\}$$

will have the property that, for a set of sequences $\{x_i\}$ of positive measure, x_1^n will belong to $\mathscr{B}(n, \varepsilon)$ for infinitely many n. Since each $\mathscr{B}(n, \varepsilon)$ can be covered by fewer than $2^{n(R-\varepsilon)}$ D-balls this would contradict Lemma 5.

A more minor observation is the following: Recall that the rate-distortion function was defined as the best one can do asymptotically on the average with block codes if an arbitrarily small part of the space is removed. Lemma 5 shows that it can be defined as the best one can do asymptotically on the average with variable-length codes on the entire space. In other words, $R_\mu(D) = \lim_{n\to\infty} R_n^*(D)$, where $R_n^*(D)$ is the greatest lower bound of the rates $R(c_n)$ over the class of *all* D-semifaithful n-codes c_n, block or variable length.

Acknowledgment. We wish to give special thanks to Imre Csiszár, who corrected several of our errors and made many suggestions for improvement of our discussion.

REFERENCES

[1] DAVISSON, L. D. (1973). Universal noiseless coding. *IEEE Trans. Inform. Theory* **IT-19** 783–795.

[2] ELIAS, P. (1975). Universal codeword sets and representations of the integers. *IEEE Trans. Inform. Theory*. **IT-21** 194–203.

[3] FELDMAN, J. (1980). r-entropy, equipartition, and Ornstein's isomorphism theorem. *Israel J. Math.* **36** 321–343.

[4] FITTINGHOF, B. M. (1966). Optimal coding in the case of unknown and changing message statistics. *Problems Inform. Transmission* **IT-2** 3–11.

[5] KIEFFER, J. (1978). A unified approach to weak universal source coding. *IEEE Trans. Inform. Theory* **IT-24** 674–682.

[6] LYNCH, T. J. (1966). Sequence time coding for data compression. *Proc. IEEE* **54** 1490–1491.

[7] ORNSTEIN, D. and WEISS, B. (1983). The Shannon–McMillan–Breiman theorem for a class of amenable groups. *Israel J. Math.* **44** 53–60.

[8] ORNSTEIN, D. and WEISS, B. (1990). How sampling reveals a process. *Ann. Probab.* **18**. To appear.

[9] SHIELDS, P. (1987). The ergodic and entropy theorems revisited. *IEEE Trans. Inform. Theory* **IT-33** 263–266.

[10] ZIV, J. (1972). Coding of sources with unknown statistics. I. Probability of encoding error. II. Distortion relative to a fidelity criterion. *IEEE Trans. Inform. Theory* **IT-18** 384–389, 389–394.

[11] ZIV, J. and LEMPEL, A. (1977). A universal algorithm for sequential data compression. *IEEE Trans. Inform. Theory* **IT-23** 337–343.

DEPARTMENT OF MATHEMATICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF TOLEDO
TOLEDO, OHIO 43606