CCIT Report #609
December 2006

On Finite Memory Universal Data Compression and Classification of Individual Sequences

Jacob Ziv
Department of Electrical Engineering
Technion–Israel Institute of Technology
Haifa 32000, Israel

November 19, 2006

Abstract

Consider the case where consecutive blocks of N letters of a semi-infinite individual sequence $\mathbf{X}$ over a finite-alphabet are being compressed into binary sequences by some one-to-one mapping. No a-priori information about $\mathbf{X}$ is available at the encoder, which must therefore adopt a universal data-compression algorithm.

It is known that if the universal LZ77 data compression algorithm is successively applied to N-blocks then the best error-free compression, for the particular individual sequence $\mathbf{X}$ is achieved as N tends to infinity.

The best possible compression that may be achieved by *any* universal data compression algorithm for *finite* N-blocks is discussed. It is demonstrated that context tree coding essentially achieves it.

Next, consider a device called *classifier* (or discriminator) that observes an individual *training* sequence $\mathbf{X}$. The classifier's task is to examine individual test sequences of length N and decide whether the test N-sequence has the same features as those that are captured by the training sequence $\mathbf{X}$, or is sufficiently different, according to some appropriate criterion. Here again, it is demonstrated that a particular universal context classifier with a storage-space complexity that is linear in N, is essentially optimal. This may contribute a theoretical "individual sequence" justification for the Probabilistic Suffix Tree (PST) approach in learning theory and in computational biology.

Index Terms: Data compression, universal compression, universal classification, context-tree coding.

A. Introduction and Summary of Results:

Traditionally, the analysis of information processing systems is based on a certain modeling of the process that generates the observed data (e.g an ergodic process). Based on this a-priori model, a processor (e.g. a compression algorithm, a classifier, etc) is then optimally designed. In practice, there are many cases where insufficient a-priori information about this generating model is available and one must base the design of the processor on the observed data only, under some complexity constraints that the processor must comply with.

1. Universal Data Compression with Limited Memory

The Kolmogorov-Chaitin complexity (1968) is the length of the shortest *program* that can generate the given individual sequence via a universal Turing machine. More concrete results are achieved by replacing the universal Turing machine model with the more restricted finite-state machine model.

The Finite- State(FS) normalized complexity (compression) $H(\mathbf{X})$, measured in bits per input letter, of an individual infinite sequence $\mathbf{X}$ is the normalized length of the shortest one-to-one mapping of $\mathbf{X}$ into a binary sequence that can be achieved by any finite-state compression device[1]. For example, the counting sequence 0123456... when mapped into the binary sequence 0,1,00,01,10,11,000,001,010,011,100,101,110,111... is incompressible by any finite-state algorithm. Fortunately, the data that one has to deal with is in many cases compressible.

The FS complexity was shown to be asymptotically achieved by applying the LZ universal data compression algorithm [1] to consecutive blocks of the individual sequence. The FS modeling approach was also applied to yield asymptotically optimal universal prediction of individual sequences[9].

Consider now the special case of a FS class of processors is further constrained to include *only* block-encoders that process one N-string at a time and then start all over again, (e.g. due to bounded latency and error-propagation considerations), $H(\mathbf{X})$ is still *asymptotically* achievable by the LZ algorithm when applied on-line to consecutive strings of length N, as N tends to infinity[1]. But the LZ algorithm may not be the best on-line universal data compression algorithm when the block-length is of *finite* length N.

It has been demonstrated that if it is a-priori known that $\mathbf{X}$ is a realization of a stationary ergodic, Variable Length Markov Chain (VLMC) that is governed by a tree model, then context-tree coding yields a smaller redundancy than the LZ algorithm([10],[4]). More recently, it has been demonstrated that context-tree coding yields an optimal universal coding policy (relative to the VLMC assumption) ([2]).

Inspired by these results, one may ask whether the optimality of context-tree coding relative to tree models still holds for more general setups.

It is demonstrated here that the best possible compression that may be achieved by *any* universal data compression algorithm for *finite* N-blocks is essentially achieved by context-tree coding for *any* individual sequence $\mathbf{X}$ and not just for individual sequences that are realizations of a VLMC.

In the following, a number of quantities are defined, that are characterized by non-traditional notations that seem unavoidable due to end-effects resulting from the finite length of X_1^N. These end-effects vanish as N tends to infinity, but must be taken into account here.

Refer to an arbitrary sequence over a finite alphabet $\mathbf{A}$, $|\mathbf{A}| = A$, $X^N = X_1^N = X_1, X_2, ...; X_N \in \mathbf{A}$, as being an *individual* sequence. Let $\mathbf{X} = X_1, X_2, ...$ denote a semi-infinite sequence over the alphabet $\mathbf{A}$. Next, an empirical probability distribution $P_{MN}(Z_1^N, N)$ is defined for N vectors that appear in a sequence of length MN. The reason for using the notation $P_{MN}(Z_1^N, N)$ rather than, say, $P_{MN}(Z_1^N)$ is due to end-effects as discussed below. We then define an empirical entropy that results from $P_{MN}(Z_1^N, N)$, namely $H_{MN}(N)$. This quantity is similar to the classical definition of the empirical entropy of N-blocks in an individual sequence of length MN and as one should anticipate, serves as a lower bound for the compression that can be achieved by any N- block encoders.

Furthermore, $H_{MN}(N)$ is achievable in the impractical case where one is allowed to first scan the long sequence X_1^{MN}, generate the corresponding empirical probability $P_{MN}(Z_1^N, N)$ for each N-vector Z_1^N that appears in X_1^{MN}, and apply the corresponding Huffman coding to consecutive N-blocks.

Then, define $H(\mathbf{X}, N) = \limsup_{M\to\infty} H_{MN}(N)$. It follows that

$$H(\mathbf{X}) = \limsup_{N\to\infty} H(\mathbf{X}, N))$$

is the smallest number of bits per letter that can be asymptotically achieved by any N-block data-compression scheme for $\mathbf{X}$. However, in practice, universal data-compression is executed on-line and the only available information on $\mathbf{X}$ is the currently processed N-block.

Next, an empirical probability measure $P_{MN}(Z_1^i, N)$ is defined for $i < N$ vectors that appear in an MN-sequence, which is derived from $P_{MN}(Z_1^N, N)$ by summing up $P_{MN}(Z_1^N, N)$ over the last $N - i$ letters of N vectors in the MN- sequence. Again, observe that due to end-effects, $P_{MN}(Z_1^i, N)$ is different from $P_{MN}(Z_1^i, i)$ but converges to it asymptotically, as M tends to infinity. Similarly, an empirical entropy $H_{MN}(i, N)$ that is derived from $P_{MN}(Z_1^i, N)$. In the analysis that follows below both $H_{MN}(N)$ and $H_{MN}(i, N)$ play an important role.

An empirical entropy $H_{MN}(Z_i|Z_1^{i-1}, N)$ is associated with each vector $Z_1^{i-1}; 1 \leq i \leq (\log N)^2$ in X_1^{MN}. This entropy is derived from $P_{MN}(Z_i|Z_1^{i-1}, N) = \frac{P_{MN}(Z_1^i, N)}{P_{MN}(Z_1^{i-1}, N)}$. Note that this empirical entropy is conditioned on the *particular* value of Z_1^{i-1} and is *not* averaged over all $Z_1^{i-1} \in \mathbf{A}^{i-1}$ relative to $P_{MN}(Z_1^{i-1}, N)$.

A context-tree with approximately N leaves, that consists of the N most empirically probable contexts in X_1^{MN}, is generated. For each leaf of this tree, choose the one context among the contexts on the the path from the root of the tree to this leaf, for which the associated entropy is the smallest. Then, these minimal associated entropies are averaged over the set of leaves of the tree. This average entropy is denoted by $H_u(N, M)$. Note that $H_u(N, M)$ is essentially an empirical conditional entropy which is derived for a suitably derived variable-length Markov chain (VLMC).

Finally, define $H_u(\mathbf{X}, N) = \limsup_{M \to \infty} H_u(N, M)$. It is demonstrated that $\liminf_{N \to \infty}[H(\mathbf{X}, N) - H_u(\mathbf{X}, N)] > 0$. Thus, for large enough N, $H_u(\mathbf{X}, N)$, like $H(\mathbf{X}, N)$, may also serve as a lower bound on the compression that may be achieved by any encoder for N-sequences. The relevance of $H_u(\mathbf{X}, N)$ becomes apparent when it is demonstrated in Theorem 2 below that a context-tree universal data-compression scheme, when applied to N'-blocks, essentially achieves $H_u(\mathbf{X}, N)$ for any $\mathbf{X}$ if $\log N'$ is only slightly larger than $\log N$, and achieves $H(\mathbf{X})$ as N" tends to infinity.

Furthermore, it is shown in Theorem 1 below that among the many compressible sequences $\mathbf{X}$ for which $H_u(\mathbf{X}, N) = H(\mathbf{X}) < \log A$, there are some for which *no* on-line universal data-compression algorithm can achieve any compression at all when applied to consecutive blocks of length N', if $\log N'$ is slightly smaller than $\log N$. Thus, context-tree universal data-compression is therefore essentially optimal. Note that the threshold effect that is described above is expressed in a logarithmic scaling of N. At the same time, the logarithmic scaling of N is apparently the natural scaling for the length of contexts in a context-tree with N leaves.

2. Application to Universal Classification

A device called *classifier* (or discriminator) observes an individual training sequence of length of m letters, X_1^m.

The classifier's task is to consider individual test sequences of length N and decide whether the test N-sequence has, in some sense the same properties as those that are captured by the training sequence, or is sufficiently different, according to some appropriate criterion. No a-priori information about the test sequences is available to the classifier asides from the training sequence.

A universal classifier $d(X_1^m, Z_1^N \in \mathbf{A}^N)$ for N-vectors is defined to be a mapping from $\mathbf{A}^N$ onto $\{0, 1\}$. Upon observing Z_1^N, the classifier declares Z_1^N to be *similar* to one of the N-vectors $X_{j+1}^{j+N}; j = 0, 1, ..., m - N$ iff $d(X_1^m, Z_1^N) = 1$ (or, in some applications, if a slightly distorted version $\tilde{Z}_1^N$ of Z_1^N satisfies $d(X_1^m, \tilde{Z}_1^N) = 1$).

In the *classical* case, the probability distribution of N-sequences is known and an optimal classifier accepts all N-sequences Z_1^N such that the probability $P(Z_1^N)$ is bigger than some preset threshold. If X_1^L is a realization of a stationary ergodic source one has, by the Asymptotic Equipartition Property (A.E.P) of information theory, that the classifier's task is tantamount (almost surely for large enough N and M) to deciding whether the test sequence is equal to a "typical" sequence of the source (or, when some distortion is acceptable, if a slightly distorted version of the test sequence is equal to a "typical" sequence). The cardinality of the set of typical sequences is, for large enough N, about 2^{NH}, where H is the entropy rate of the source[10].

What to do when $P(Z_1^N)$ is unknown or does not exist, and the only available information about the generating source is a training sequence X_1^m? The case where the training sequence is a realization of an ergodic source with vanishing memory is studied in [11], where it demonstrated that a certain universal context-tree based classifier is essentially optimal for this class of sources. This is in unison with related results on universal prediction([11],[12],[13]).

Universal classification of test sequences relative to a long training sequence is a central problem in computational biology. One common approach is to assume that the training sequence is a realization of a VLMC, and upon viewing the training sequence, to construct an empirical Probabilistic Suffix Tree (PST), the size of which is limited to the available storage complexity of the classifier, and apply a context-tree based classification algorithm [7,8].

But how one should proceed if there is no a-priori support for the VLMC assumption? In the following, it is demonstrated that the PST approach is essentially optimal for every *individual* training sequence, even without the VLMC assumption. Denote by $S_\epsilon(N, X_1^m)$ a set of N-sequences Z_1^N which are declared to be similar to X_1^m (i.e. $d(X_1^m, Z_1^N) = 1$), where $d(X_1^m, X_{j+1}^{j+N}) = 0$ should be satisfied by no more than $\epsilon(m - N + 1)$ instances $j = 0, 1, 2, ..., m - N$, and where ϵ is an arbitrarily small positive number. Also, given a particular classifier, let $D_\epsilon(N, X_1^m) = |S_\epsilon(N, X_1^m)|$, and let

$$H_\epsilon(N, X_1^m) = \frac{1}{N} \log D_\epsilon(N, X_1^m) \,.$$

Thus, any classifier is characterized by a certain $H_\epsilon(N, X_1^m)$. Given X_1^m, let $D_{\epsilon,min}(N, X_1^m)$ be the smallest achievable $D_\epsilon(N, X_1^m)$ and let

$$H\epsilon, min(N, X_1^m) = \frac{1}{N} \log D_{\epsilon,min}(N, X_1^m) \,.$$

and, for an infinite training sequence $\mathbf{X}$,

$$H_\epsilon(N, \mathbf{X}) = \limsup_{m \to \infty} H_{\epsilon,min}(N, X_1^m)$$

.

Note that

$$\bar{H}(\mathbf{X}) = \lim_{\epsilon \to 0} \limsup_{N \to \infty} H_{\epsilon,min}(N, \mathbf{X})$$

is the topological entropy of $\mathbf{X}$[6].

Naturally, if the classifier has the complete list of N-vectors that achieve $D_{\epsilon,min}(N, X_1^m)$, it can achieve a perfect classification by making $d(X_1^m, Z_1^N) = 1$ iff $Z_1^N = X_{j+1}^{j+N}$, for every instant $j = 0, 1, 2, ..., m - N$ for which $X_{j+1}^{j+N} \in S(N, \epsilon, X_1^m)$.

The discussion is constrained to cases where $H\epsilon, min(N, X_1^m) > 0$. Therefore, when m is large, $D_{\epsilon,min}(N, X_1^m)$ grows exponentially with N(e.g. when the test sequence is a realization of an ergodic source with a positive entropy rate). The attention is limited to classifiers that have a storage-space complexity that grows only linearly with N. Thus, the long training sequence cannot be stored. Rather, the classifier is constrained to represent the long training sequence by a short "signature", and use this short signature to classify incoming test sequences of length N. It is shown that it is possible to find such a classifier, denoted by $d(X_1^m, \epsilon, Z_1^N)$, which is essentially optimal in the following sense:

An optimal "ϵ-efficient" universal classifier $d(X_1^m, \epsilon, Z_1^N)$ is defined to be one that satisfies the condition that $d(X_1^m, X_{j+1}^{j+N}) = 1$ for $(1 - \hat{\epsilon})(m - N + 1)$ instances $j = 0, 1, ...m - N$, where $\hat{\epsilon} \leq \epsilon$. This corresponds to the rejection of at most $\epsilon D_{\epsilon,min}(N, X_1^m)$ vectors from among the $D_{\epsilon,min}(N, X_1^m)$ *typical* N-vectors in X_1^m. Also, an optimal "ϵ-efficient" universal classifier should satisfy the condition that $d(X_1^m, Z_1^N) = 1$ is satisfied by no more than $2^{NH_{\epsilon,min}(N,X_1^m)+\epsilon}$ N-vectors Z_1^N. This corresponds to a false-alarm rate of

$$\frac{2^{N(H_{\epsilon,min}(N,X_1^m)+\epsilon)} - 2^{NH_{\epsilon,min}(N,X_1^m)}}{2^{N \log A} - 2^{NH_{\epsilon,min}(N,X_1^m)}}$$

when N-vectors are selected independently at random, with an induced uniform probability distribution over the set of $2^{N \log A} - 2^{N(H_{\epsilon,min}(N,X_1^m))}$ N-vectors that should be rejected. Note that the false-alarm rate is thus guaranteed to decrease exponentially with N for any individual sequence X_1^m for which $H_{\epsilon,min}(N, X_1^m) < \log A - \epsilon$.

A context-tree based classifier for N-sequences, given an infinite training sequence $\mathbf{X}$ and a storage-complexity of $O(N)$, is shown by Theorem 3 below to be ϵ-efficient for any $N \geq N_0(\mathbf{X})$ and some $m = m_0(N, \mathbf{X})$. Furthermore, by Theorem 3 below, among the set of training sequences for which the proposed classifier is ϵ-efficient, there are some for which no ϵ-efficient classifier for N'-sequences exists, if $\log N' < \log N$ for any $\epsilon < \log A - H_{\epsilon,min}(N, \mathbf{X})$. Thus, the proposed classifier is essentially optimal.

Finally, the following universal classification problem is considered: Given two test-sequences Y_1^N and Z_1^N and *no* training data, are these two test-sequences "similar" to each other? The case where both Y_1^N and Z_1^N are realizations of some (unknown) finite-order Markov processes is discussed in [14], where an asymptotically optimal empirical divergence measure is derived empirically from Y_1^N and Z_1^N.

In the context of the individual-sequence approach that is adopted here, this amounts to the following problem: Given Y_1^N and Z_1^N, is there a training-sequence $\mathbf{X}$ for which $H_{\epsilon,min}(\mathbf{X}) > 0$,such that both Y_1^N and Z_1^N are accepted by some ϵ-efficient universal classifier with linear space complexity?(this problem is a reminiscence of the "common ancestor" problem in computational biology[15], where one may think of $\mathbf{X}$ as a training sequence that captures the properties of a possible "common ancestor" of two DNA sequences Y_1^N and Z_1^N).

This is the topic of the Corollary following Theorem 3 below.

B. Definitions,Theorems and Algorithms

Given $X_1^N \in \mathbf{A}^N$, let $c(X_1^N); X \in \mathbf{A}^N$ be a one-to-one mapping of X_1^N into a binary sequence of length $L(X_1^N)$, which is called the length function of $c(X_1^N)$. It is assumed that $L(X_1^N)$ satisfies the Kraft inequality.

For every $\mathbf{X}$ and any positive integers M, N, define the compression of the prefix X_1^{NM} to be:

$$\rho_L(\mathbf{X}, N, M) = \max_{i;1\le i\le N-1} \frac{1}{NM}\left[\left[\sum_{j=0}^{M-2} L\Big(X_{i+jN+1}^{(i+(j+1)N)}\Big)\right] + L(X_1^i) + L\Big(X_{(i+1+(M-1)N)}^{NM}\Big)\right].$$

Thus, one looks for the compression of the sequence X_1^{MN} that is achieved by successively applying a given length-function $L(X_1^N)$ and with the *worst* starting *phase* $i; i = 1, 2, ..., N-1$. Observe that by ignoring the terms $\frac{1}{NM}L(X_1^i)$ and $\frac{1}{NM}L\Big(X_{(i+1+(M-1)N)}^{NM}\Big)$ (that vanish for large values of M) in the expression above, one gets a lower bound on the actual compression.

In the following, a lower-bound $H_{MN}(N)$ on $\rho_L(\mathbf{X}, N, M)$ is derived, that applies to any length-function $L(X_1^N)$. First, the notion of *empirical probability* of an N-vector in a *finite* MN-vector is derived, for any two positive integer N and M.

Given a sequence X_1^{MN}, define for a vector $Z_1^N \in \mathbf{A}^N$,

$$P_{i,MN}(Z_1^N, N) = \frac{1}{M-1}\sum_{j=0}^{M-2} \mathbb{1}_{Z_1^N}\left(X_{i+jN}^{i+(j+1)N-1}\right); 1 \le i \le N-1 \tag{1}$$

and

$$P_{MN}(Z_1^N, N) = \frac{1}{N}\sum_{i=1}^{n} P_{i,MN}(Z_1^N, N) \tag{2}$$

where,

$$\mathbb{1}_{Z_1^N}(X_{jN+i}^{(j+1)N+i-1}) = 1 \text{ iff } X_{jN+i}^{(j+1)N+i-1} = Z_1^N; \text{ else } \mathbb{1}_{Z_1^N}(X_{jN+i}^{(j+1)N+i-1}) = 0.$$

Thus,

$$P_{MN}(Z_1^N, N) = \frac{1}{(M-1)N+1} \sum_{i=1}^{(M-1)N+1} \mathbb{1}_{Z_1^N}(X_i^{i+N-1}),$$

the empirical probability of Z_1^N.

Similarly, define

$$P_{MN}(Z_1^i, N) = \sum_{Z_{N-i+1}^N \in \mathbf{A}^i} P_{MN}(Z_1^N, N); 1 \le i \le N-1$$

(As noted in the Introduction, $P_{MN}(Z_1^i, N)$ converges to the empirical probability $P_{MN}(Z_1^i, i)$ as M tends to infinity. However, for finite values of M, these two quantities are not identical due to end-effects).

Let,

$$H_{MN}(i, N) = -\frac{1}{\ell} \sum_{Z_1^i \in \mathbf{A}^i} P_{MN}(Z_1^i, N) \log P_{MN}(Z_1^i, N)$$

and

$$H_{MN}(N) = H_{MN}(N, N) = -\frac{1}{N} \sum_{Z_1^N \in \mathbf{A}^N} P_{MN}(Z_1^N, N) \log P_{MN}(Z_1^N, N) \tag{3}$$

then,

Proposition 1

$$\rho_L(\mathbf{X}, N, M) \ge H_{MN}(N)$$

and

$$\limsup_{M\to\infty} \rho_L(\mathbf{X}, N, M) \ge H(\mathbf{X}, N)$$

where $H(\mathbf{X}, N) = \limsup_{M\to\infty} H_{MN}(N)$. The proof appears in the Appendix.

Thus, the best possible compression of X_1^{MN} that may be achieved by any one-to-one encoder for N-blocks, is bounded from below by $H_{MN}(N)$. Furthermore, $H_{MN}(N)$ is achievable for N that is much smaller than $\log M$, if $c(Z_1^N)$ (and its corresponding length-function $L(Z_1^N)$ is tailored to the individual sequence X_1^{MN}, by first scanning X_1^{MN}, evaluating the empirical distribution $P_{MN}(Z_1^N, N)$ of N-vectors and then applying the corresponding Huffman data compression algorithm. However, in practice, the data-compression has to be executed on-line and the only available information on X_1^{MN} is the one that is contained the currently processed N-block.

The main topic of this paper is to find out how well one can do in such a case where the *same* mapping $c(X_1^N)$ is being applied to successive N-vectors of X_1^{MN}.

Next, given N and M a particular context-tree is generated from X_1^{MN} for each letter $X_i; 1 \leq i \leq MN$, and define a related conditional empirical entropy $H_u(N, M)$ that corresponds to these context trees. It is then demonstrated that for large enough N and M, $H_u(N, M)$ may also serve as a lower bound on $\rho_L(\mathbf{X}, N, M)$.

Construction of the Context-tree for the letter Z_i

1) Consider contexts which are no longer than $t = \lceil (\log N)^2 \rceil$ and let K be a positive number.

2)Let $K_1(Z_1^N, K) = \min[j-1, t]$ where j is the smallest positive integer such that $P_{MN}(Z_1^j, N) \leq \frac{1}{K}$, where the probaility measure $P_{MN}(Z_1^j, N)$ for vectors $Z_1^j \in \mathbf{A}^j$ is dervided from X_1^{MN}. If such j does not exist, set $K_1(Z_1^N, K) = -1$, where Z_1^0 is the null vector.

3)Given X_1^{MN} evaluate $P_{MN}(Z_1^i, N)$. For the i-th symbol in Z_1^N, let Z_1^{i-1} be the corresponding suffix. For each particular $Z_1^{i-1} \in \mathbf{A}^{i-1}$ define

$$H_{MN}(Z_i|Z_1^{i-1}, N) = -\sum_{Z_i \in \mathbf{A}} \frac{P_{MN}(Z_1^i, N)}{P_{MN}(Z_1^{i-1}, N)} \log \frac{P_{MN}(Z_1^i, N)}{P_{MN}(Z_1^{i-1}, N)} \tag{4}$$

4)Let $j_0 = j_0(Z_1^{i-1})$ be the integer for which,

$$H_{MN}(Z_i|Z_{i-j_0}^{i-1}, N) = \min_{1 \leq j \leq 1 + K_1(Z_1^N, K)} H_{MN}(Z_i|Z_{i-j}^{i-1}, N)$$

Each such j_0 is a node in a tree with about K leaves. The set of all such nodes represent the particular context tree for the i-th instant.

5)Finally,

$$H_u(N, K, M) = \sum_{Z_1^N \in \mathbf{A}^N} P_{MN}(Z_1^N, N) H_{MN}(Z_i|X_{i-j_0(Z_1^{i-1})}^{i-1}, N) \tag{5}$$

Observe that $H_u(N, K, M)$ is an entropy-like quantity defined by an optimal data-driven tree of variable depth K_1, where each leaf that is shorter than t has roughly an empirical probability $\frac{1}{K}$.

Set $K = N$ and let $H_u(N, M) = H_u(N, N, M)$. Also, let

$$H_u(\mathbf{X}, N) = \limsup_{M \to \infty} H_u(N, M) \tag{6}$$

and

$$H_u(\mathbf{X}) = \limsup_{N \to \infty} H_u(\mathbf{X}, N). \tag{7}$$

Let

$$H(\mathbf{X},N) = \limsup_{M\to\infty} H_{MN}(N)$$

and

$$H(\mathbf{X}) = \limsup_{N\to\infty} H(\mathbf{X},N)$$

Then,

Lemma 1 *For every individual sequence X,*

$$\liminf_{N\to\infty} \Big[H(\mathbf{X},N) - H_u(\mathbf{X},N) \Big] \geq 0 \tag{8}$$

Hence, $H_u(\mathbf{X}) \leq H(\mathbf{X})$. The proof of Lemma 1 appears in the Appendix.

A compression algorithm that achieves a compression $H_u(\mathbf{X},N)$ is therefore *asymptotically* optimal as N tends to infinity.

Note that the conditional entropy for a data-driven Markov tree of a uniform depth of, say, $O(\log\log N)$ may still satisfy Lemma 1 (by the proof of Lemma 1), but this conditional empirical entropy is lower-bounded by $H_u(\mathbf{X},N)$ for *finite* values of N.

A context-tree data-compression algorithm for N-blocks that essentially achieves a compression $H_u(\mathbf{X},N)$ is introduced below, and is therefore asymptotically optimal, but so are other universal data-compression algorithms (e.g., a simpler context-tree data compression algorithm with a uniform context depth or the LZ algorithm [1]). However, the particular context-tree algorithm that is proposed below is shown to be essentially optimal for *non-asymptotic* values of N as well.

It is now demonstrated that no universal data-compression algorithm that utilizes a length-function for N-blocks can always achieve a compression which is essentially better than $H_u(\mathbf{X},N)$ for *any* value of N in the following sense:

Let δ be an arbitrarily small positive number. Let us consider the class $C_{N_0,M_0,\delta}$ of all $\mathbf{X}$-sequences for which for some $\hat{H}$ such that $\delta < \hat{H} < (1-2\delta)\log A$,

1) $H_{M_0N_0}(N_0,N_0) = \hat{H}$.
2) $H_u(N_0,K_0,M_0) - \hat{H} \leq \delta$ where $K_0 = N_0$.

It is demonstrated that the class $C_{N_0,M_0,\delta}$ is not empty. In the proof of Theorem 1 in the Appendix, a class of cardinality $M_{\ell,h} = \binom{2^{(\log A)\ell}}{2^{h\ell}}$ of sequences is constructed such that this class is included in $C_{N_0,M_0,\delta}$ for $h = \frac{\delta}{2}$ and for ℓ that satisfy $N_0 = 2^{h\ell}$. Moreover, by Lemma 1, it follows that every sequence $\mathbf{X}$ is in the set $C_{N_0,M_0,\delta}$, for large enough N_0 and $M_0 = M_0(N_0)$.

Theorem 1 *Let* $N' = N^{1-\delta}$. *For any universal data-compression algorithm for* N'*-vectors that utilizes some length-function* $L(Z_1^{N'})$, *there exist some sequences* $\mathbf{X} \in C_{N_0,M_0,\delta}$ *such that for any* $M \geq M_0$ *and any* $N \geq N_0$:

$$\rho_L(\mathbf{X}, M, N') \geq (1-\delta)[\log A - \delta] > \hat{H}$$

for large enough N_0.

The proof of Theorem 1 appears in the Appendix.

The next step demonstrates that there exists a universal data-compression algorithm, which is optimal in the sense that when it is applied to consecutive N-blocks, its associated compression is about $H_u(\mathbf{X}, N')$ for *every* individual sequence $\mathbf{X}$ where $\log N'$ is slightly smaller than $\log N$.

Theorem 2 *Let* δ *be an arbitrarily small positive number and let* $N' = \lfloor N^{1-\delta} \rfloor$. *There exists a context-tree universal coding algorithm for* N*-blocks, with a length-function* $\hat{L}(Z_1^N)$ *for which, for every individual* $X_1^{MN} \in \mathbf{A}^{MN}$,

$$\rho_{\hat{L}}(\mathbf{X}, M, N) \leq H_u(N, N', M) + O\left(\frac{\log N}{N^\delta}\right)$$

It should be noted here that the particular universal context-tree algorithm that is described below is not claimed to yield the lowest possible redundancy for a given block-length N. No attempt was taken to minimize the redundancy, since it is suffice to establish the fact that this particular *essentialy optimal* universal algorithm indeed belongs to the class of context-tree algorithms. The reader s referred to [4] for an exhaustive discussion of optimal universal context-tree algorithms.

Description of the universal compression algorithm: Consider first the encoding of the first N-vector X_1^N (to be repeated for every $X_{(i-1)N+1}^{iN}; i = 2, 3, ..., M-1$).

Let $t = \lceil (\log N)^2 \rceil$.

A) Given the first N-vector X_1^N, generate the set $T_u(X_1^N)$ that consists of all contexts Z_1^{i-1} that appear in X_1^N, satisfying $P_N(Z_1^{i-1}, t) \geq \frac{1}{N^{1-\delta}}; i \leq t$.

Clearly, $T_u(X_1^N)$ is a context tree with no more than $N^{1-\delta}$ leaves with a maximum depth of $t = \lceil (\log N)^2 \rceil$. The depth t is chosen to be just small enough so as to yield an *implementable* compression scheme and at the same time, still be big enough so as to yield an *efficient* enough compression.

B) Evaluate,

$$H_u(X_1^N, T_u, t) = \sum_{X_1^{t-1} \in \mathbf{A}^{t-1}} P_N(X_1^{t-1}, t) \min_{0 \le j \le t-1; X_1^{j-1} \in T_u(X_1^N)} H_N(X_i | X_1^{i-1}, t) \,.$$

Let $\hat{T}_u(X_1^N)$ be a sub-tree of $T_u(X_1^N)$, such that it's leaves are the the set of contexts that achieves $H_u(X_1^N, T_u, t)$.

C) A length function $\hat{L}(X_1^N) = \hat{L}_1(X_1^N) + \hat{L}_2(X_1^N) + \hat{L}_3(X_1^N)$ is constructed as follows:

1) $\hat{L}_1(X_1^N)$ is the length of an uncompressed binary length-function, $\hat{m}_1(X_1^N)$ that enables the decoder to reconstruct the context tree $\hat{T}_u(X_1^N)$, that consists of the set of contexts that achieves $H_u(X_1^N, T_u, t)$. This tree has, by construction, at most $N^{1-\delta}$ leaves and at most t letters per leaf. It takes at most $1 + t\log A$ bits to encode a vector of length t over an alphabet of A letters. It also takes at most $1 + \log t$ bits to encode the length of a particular context. Therefore,
$\hat{L}_1(X_1^N) \le N^{1-\delta}(t \log A + \log t + 2) \le [\log N \log A + \log(\log N^2) + 2]N^{1-\delta}$ bits.

2) $\hat{L}_2(X_1^N)$ is the length of a binary word $\hat{m}_2(X_1^N)$, ($t \log A$ bits long), which is an uncompressed binary mapping of X_1^t, the first t letters of X_1^N.

3) Observe that given $\hat{m}_1(X_1^N)$, and $\hat{m}_2(X_1^N)$, the decoder can re-generate X_1^t and the sub-tree $\hat{T}_u(X_1^N)$ that achieves $H_u(X_1^N, T_u, t)$). Given $\hat{T}_u(X_1^N)$ and a prefix X_1^t of X_1^N, X_{t+1}^N is compressed by a context-tree algorithm for FSMX sources [3,4], which is tailored to the contexts that are the leaves of $\hat{T}_u(X_1^N)$, yielding a length function $\hat{L}_3(X_1^N) \le NH_u(X_1^N, T, t) + O(1)$.

D) Repeat the steps 1), 2) and 3) above for the N-vectors $X_{(i-1)N+1}^{iN}; i = 2, 3, ..., M-1$.

Let $\bar{T}_u(X_1^N)$ be the set of all vectors Z_1^{i-1} satisfying $P_{MN}(Z_1^{i-1}, N) \ge \frac{1}{N"}; i \le t$, where $N" = N^{1-2\delta}$.

The proof of Theorem 2 follows from the construction and by the convexity of the entropy function since

$$\frac{1}{M} \sum_{i=0}^{M-1} H_u\left(Z_{iN+1}^{(i+1)N}, T_u, t\right) \le H_u(N, N", M) + O(N^{-\delta}) + O\left(\frac{[\log N]^4}{N}\right)$$

and where the term $O(N^{-\delta})$ is an upper-bound on the relative frequency of instances in $X_{iN+1}^{(i+1)N}$ that have as a context a leaf of $\hat{T}_u(X_{iN+1}^{(i+1)N})$ that is a suffix of one of the vectors in $\bar{T}_u(X_1^N)$ and therefore is not an element of the set of contexts that achieve $H_u(N, N", M)$. The term $O(\frac{[\log N]^4}{N})$ is due to end-effects and follows from Lemma 2.7 in [5,page 33](see proof of Lemma 1 in the Appendix).

C. Application to Universal Classification

A device called *classifier* (or discriminator) observes an individual training sequence of length of m letters, X_1^m.

The classifier's task is to consider individual test sequences of length N and decide whether the test N-sequence has the same features as those that are captured by the training sequence, or is sufficiently different, according to some appropriate criterion. No a-priori information about the test sequences is available to the classifier aside from the training sequence. Following the discussion in the Introduction section, a universal classifier $d(X_1^m, Z_1^N \in \mathbf{A}^N)$ for N-vectors is defined to be a mapping from $\mathbf{A}^N$ onto $\{0,1\}$. Upon observing Z_1^N, the classifier declares Z_1^N to be *similar* to one of the N-vectors $X_{j+1}^{j+N}; j = 0, 1, ..., m-N$ iff $d(X_1^m, Z_1^N) = 1$ (or, in some applications, if a slightly distorted version $\tilde{Z}_1^N$ of Z_1^N satisfies $d(X_1^m, \tilde{Z}_1^N) = 1$). Denote by $S(N, \epsilon, X_1^m)$ a set of N-sequences Z_1^N which are declared to be similar to X_1^m, i.e. $d(X_1^m, Z_1^N) = 1$), where $d(X_1^m, X_{j+1}^{j+N}) = 0$ should be satisfied by no more than $\epsilon(m - N + 1)$ instances $j = 0, 1, 2, ..., m - N$, and where ϵ is an arbitrarily small positive number. Also, given a particular classifier, let $D_\epsilon(N, X_1^m) = |S(N, \epsilon, X_1^m)|$, and let

$$H_\epsilon(N, X_1^m) = \frac{1}{N} \log D_\epsilon(N, X_1^m) \,.$$

Thus, any classifier is characterized by a certain $H_\epsilon(N, X_1^m)$. Given X_1^m, let $D_{\epsilon,min}(N, \epsilon, X_1^m)$ be the smallest achievable $D_\epsilon(N, X_1^m)$ and let

$$H\epsilon, min(N, X_1^m) = \frac{1}{N} \log D_{\epsilon,min}(N, X_1^m) \,.$$

Naturally, if the classifier has the complete list of N-vectors that achieve $D_{\epsilon,min}(N, X_1^m)$, it can perform a perfect classification by making $d(X_1^m, Z_1^N) = 1$ iff $Z_1^N = X_{j+1}^{j+N}$ for every instant $j = 0, 1, 2, ..., m - N$ for which $X_{j+1}^{j+N} \in S(N, \epsilon, X_1^m)$. The discussion is constrained to cases where $H\epsilon, min(N, X_1^m) > 0$. Therefore, when m is large, $D_{\epsilon,min}(N, X_1^m)$ grows exponentially with N.

The attention s limited to classifiers that has a storage-space complexity that grows only linearly with N. Thus the training sequence cannot be stored. Rather, the classifier should represent the long training sequence with a short "signature" and use it to classify incoming test sequences of length N. It is shown that it is possible to find such a classifier, denoted by $d(X_1^m, \epsilon, Z_1^N)$, that is essentially optimal in the following sense(as discussed and motivated in the Introductory section): $d(X_1^m, \epsilon, Z_1^N)$ is defined to be one that satisfies the condition that $d(X_1^m, X_{j+1}^{j+N}) = 1$ for $(1 - \hat{\epsilon})(m - N + 1)$ instances $j = 0, 1, ...m - N$, where $\hat{\epsilon} \leq \epsilon$. This corresponds to a rejection of at most ϵN vectors among N-vectors in X_1^m. Also, an optimal "ϵ-efficient" universal classifier should satisfy the condition that $d(X_1^m, Z_1^N) = 1$ is satisfied by no more than $2^{NH_{\epsilon,min}(N, X_1^m)+\epsilon}$ N-vectors Z_1^N.

Observe that in the case where $\mathbf{X}$ is a realization of a finite-alphabet stationary ergodic process, $\lim_{\epsilon\to 0}\lim_{N\to\infty}\limsup_{m\to\infty} H_0(N, X_1^m)$ is equal almost surely to the entropy- rate of the source and, for a large enough N, the classifier efficiently identifies typical N-vectors without searching the exponentially large list of typical N-vectors, by replacing the long training sequence with an "optimal sufficient statistics" that occupies a memory of $O(N)$ only.

In the following, a universal context classifier for N-vectors with a a storage-space complexity that is linear in N, is shown to be essentially optimal for large enough N and m.

Description of the universal classification algorithm:

A) Evaluate $H_{\epsilon,min}(N, X_1^m)$.

B) Let $M = \lfloor \frac{m}{N} \rfloor$ and let $N" = \lfloor N^{1-2\epsilon} \rfloor$. Compute $H_u(N, N", M)$ where $H_u(N, N", M)$ is given by Eq.(5), with $N"$ replacing K. Let $\bar{T}_u(X_1^m)$ be the subset of contexts for which the minimization that yields $H_u(N, N", M)$ is achieved.

   Note that steps A), and B) above are preliminary pre-processing steps that are carried out prior to the construction of the classifier that is tailored to the training data X_1^m.

C) Compute $h_u(Z_1^N, X_1^m, \bar{T}_u, t)$

$$= - \sum_{Z_{-i+1}^0 \in \bar{T}_u(X_1^m)} P_N(Z_{-i+1}^0, t) \sum_{Z_1 \in \mathbf{A}} P_N(Z_1|Z_{-i+1}^0, t) \log P_m(Z_1|Z_{-i+1}^0, N)$$

D) Similar to step B in the description of the universal data compression algorithm above, let $T_u(X_1^N)$ be the context tree that consists of all contexts Z_1^{i-1} that appear in Z_1^N and that satisfy $P_N(Z_1^{i-1}, t) \geq \frac{1}{N^{1-\epsilon}}; i \leq t$. Let $S(Z_1^N, \mu)$ be the set of all $X_1^N \in \mathbf{A}^N$ such that $g(Z_1^N, X_1^N) \leq \mu$, where $g(*, *)$ is some non-negative distortion function satisfying: $g(Z_1^N, X_1^N) = 0$ iff $X_1^N = Z_1^N$. Given a test sequence Z_1^N, compute $H_u(\tilde{Z}_1^N, T_u, t) =$

$$\min_{\bar{Z}_1^N \in S(Z_1^N, \mu)} [- \sum_{\bar{Z}_{-i+1}^0 \in T_u(X_1^m)} P_N(\bar{Z}_{-i+1}^0, t) \sum_{\bar{Z}_1 \in \mathbf{A}} P_N(\bar{Z}_1|\bar{Z}_{-i+1}^0, t) \log P_N(\bar{Z}_1|\bar{Z}_{-i+1}^0, t)]$$

E) Let

$$\Delta(Z_1^m, Z_1^N) = h_u(Z_1^N, X_1^m, \bar{T}_u, t) - \min\Big[H_u(Z_1^N, T_u, t), H_{\epsilon,min}(N, X_1^m)\Big]$$

   and set $\hat{d}(X_1^m, \epsilon, Z_1^N \in \mathbf{A}^N)$=1 iff $\Delta(X_1^m, Z_1^N) \leq \epsilon'$, where ϵ' is set so as to guarantee that $\hat{d}(X_1^m, \epsilon, \tilde{Z}_{j+1}^{j+N}) = 1$ for some $\tilde{Z}_1^N \in S(Z_1^N, \mu)$, for at least $(1-\epsilon)(m-N+1)$ instances $j = 0, 1...., m-N$. If $H_u(Z_1^N, T_u, t) + \epsilon' > \log A$, set $\hat{d}(X_1^m, Z_1^N \in \mathbf{A}^N)$=1 for every $Z_1^N \in \mathbf{A}^N$.

Refer to a test sequence Z_1^N as being ϵ acceptable (relative to X_1^m) iff $\Delta(X_1^m, Z_1^N) \leq \epsilon'$.

It should be noted that for most instances in Z_1^N, except for at most $NO(N^{-\epsilon})$ instances in Z_1^N, a context that is an element $\bar{T}_u(X_1^m)$ is a suffix of an element in $T_u(Z_1^N)$. Hence, by the convexity of the logarithmic function, it follows that $\Delta(\mathbf{X}, Z_1^N) \geq -O(N^{-\epsilon})$. If, for example, $\mathbf{X}$ is a realization of a stationary i.i.d. process, and if no distortion is allowed $(\mu = 0)$, $\Delta(\mathbf{X}, Z_1^N) + O(N^{-\epsilon})$ is almost surely larger than or equal to the divergence

$$\sum_{Z \in \mathbf{A}} Q_{Z_1^n}(Z) \log \frac{Q_{Z_1^n}(Z)}{P(X = Z)}$$

where $P(X)$ is the probability distribution of the i.i.d process and where $Q_{Z_1^n}(Z)$ is the empirical probability of $Z \in Z_1^N$.

It should be noted that for some small values of N, one may find some values of m for which $\hat{H}_\epsilon(N, Z_1^m)$ is much larger than $H_{\epsilon,min}(\mathbf{X})$. It should also be noted that if no distortion is allowed (i.e. $\mu = 0$), the time complexity of the proposed algorithm is linear in N as well.

Theorem 3 *1)For any arbitrarily small positive ϵ, the "ϵ-efficient" classifier that is described above accepts no more than $2^{\hat{H}_\epsilon(N, Z_1^m)}$ N-vectors where*

$$\limsup_{N \to \infty} \liminf_{m \to \infty} \hat{H}_\epsilon(N, Z_1^m) \leq H_{\epsilon,min}(\mathbf{X}) + \epsilon^2$$

2) There exist m-sequences such that $\hat{H}_\epsilon(N, X_1^m)$ is much smaller than $\log A$ for which no classifier can achieve $\hat{H}_{\epsilon,min}(N', X_1^m) < \log A - \delta$ if $\log N' < \log N$, where δ is an arbitrarily small positive number.

Thus, for every $\mathbf{X}$ there exists some $N_0(\mathbf{X})$ such that for every $N \geq N_0$, the the proposed algorithm is essentially optimal for some $m_0 = m_0(N, \mathbf{X})$ and is characterized by a storage-space complexity that is linear in N. Furthermore, if one sets $\mu = 0$ (i.e. no distortion), the proposed algorithm is also characterized by a linear time-complexity. The proof of Theorem 3 appears in the Appendix. Also, it follows from the proof of Theorem 3 that if one generates a training sequence $\mathbf{X}$ such that $\liminf_{M \to \infty} H_u(N, N, M) = \lim_{M \to \infty} H_u(N, N, M)$ (i.e. a "stationary" training sequence), then there always exist positive integers N_0 and $m_0 = m_0(N_0)$ such that the proposed classifier is *essentially* optimal for *any* $N > N_0(\mathbf{X})$ and any $m > m_0(N_0)$, rather than only some specific values of m that depend on N.

Now, let Y_1^N and Z_1^N be two N-sequences and assume that no training sequence is available. However, one still would like to test the hypothesis that there exists some test sequence $\mathbf{X}$ such that both N-sequences are ϵ-acceptable with respect to $\mathbf{X}$. (This is a reminiscence of the "common ancestor" problem in computational biology where one may think of $\mathbf{X}$ as a training sequence that captures the properties of a possible "common ancestor"[15] of two DNA sequences Y_1^N and Z_1^N).

Corollary 1 *Let Y_1^N and Z_1^N be two N-sequences and let $S(Y_1^N, Z_1^N)$ be the union of all their corresponding contexts that are no longer than $t = \lceil (\log N)^2 \rceil$ with an empirical probability of at least $\frac{1}{N^{1-\delta}}$.*

If there does not exist a conditional probability distribution

$$P(X_1|X_{-i+1}^0); X_{-i+1}^0 \in S(Y_1^N, Z_1^N)$$

such that,

$$\sum_{X_1 \in \mathbf{A}} P_{N,Y_1^N}(X_1|X_{-i+1}^0, t) \log \frac{P_{N,Y_1^N}(X_1|X_{-i+1}^0)}{P(X_1|X_{-i+1}^0)} \leq \epsilon$$

and at the same time,

$$\sum_{X_1 \in \mathbf{A}} P_{N,Z_1^N}(X_1|X_{-i+1}^0, t) \log \frac{P_{N,Z_1^N}(X_1|X_{-i+1}^0, t)}{P(X_1|X_{-i+1}^0)} \leq \epsilon$$

(where $P_{N,Y_1^N}(Y_1|Y_{-i+1}^0, t)$ is empirically derived from Y_1^N and $P_{N,Z_1^N}(X_1|X_{-i+1}^0, t)$ is empirically derived from Z_1^N), then there does not exist a training sequence $\mathbf{X}$, such that for some X_1^m, $H_0(\mathbf{X}) > 0$ and $\hat{H}_\epsilon(N, X_1^m) \leq H_{\epsilon,min}(\mathbf{X}) + \epsilon^2$ (i.e. N is "long enough" relative to $\mathbf{X}$), both Y_1^N and Z_1^N are ϵ-acceptable relative to X_1^m.

In unison with the "individual sequence" justification for the essential optimality of Context-tree universal data compression algorithms [3, 4] that was established above, these results may contribute a theoretical "individual sequence" justification for the Probabilistic Suffix Tree approach in learning and in computational biology [7, 8].

Acknowledgement

Helpful discussions with Neri Merhav and helpful comments by the anonymous reviewers are acknowledged with thanks.

Appendix

Proof of Proposition 1: By definition,

$$\rho_L(\mathbf{X}, N, M) \geq \max_{i=1}^{N-1} \frac{1}{NM} \left[\sum_{j=0}^{M-2} L(X_{i+jN}^{i+(j+1)N-1}) \right] \geq \frac{1}{N} \sum_{i=1}^{N} \sum_{Z_1^N) \in \mathbf{A}^N} P_{i,MN}(Z_1^N), N) L(Z_1^N)$$

$$= \frac{1}{N} \sum_{Z_1^N \in \mathbf{A}^N} P_{MN}(Z_1^N, N) L(Z_1^N) \geq H_{MN}(N) \quad (9)$$

which leads to Proposition 1 by the Kraft inequality.

Proof of Lemma 1: Let N_0, M_0 and M be positive numbers and let $\epsilon = \epsilon(M)$ be an arbitrarily small positive number, satisfying $\log M_0 > N_0 \log A$, $H_u(\mathbf{X}) \geq H_u(\mathbf{X}, N_0) - \epsilon$, and $M \geq {N_0}^2$ such that $H_u(M_0 N_0, M_0 N_0, M) \geq H_u(\mathbf{X}) - \epsilon$, where $H_u(M_0 N_0, M_0 N_0, M) = H_u(M_0 N_0, K, M)$ with $K = M_0 N_0$.

Therefore, by the properties of the entropy function, by applying the chain-rule to $H_{MN_0}(N_0, M_0 N_0)$ and by Eq. (5),

$$H_{MM_0N_0}(N_0, M_0 N_0) \geq H_{MM_0N_0}(Z_N | Z_1^{N_0-1}, M_0 N_0) \geq H_u(M_0 N_0, M_0 N_0, M) - \frac{\log A}{N_0}$$

$$\geq H_u(\mathbf{X}, N_0) - 2\epsilon - \frac{\log A}{N_0}$$

where by Eq (4), the term $\frac{\log A}{N_0}$ is an upper-bound on the total contribution to $H_u(M_0 N_0, M_0 N_0, M)$ by vectors $Z_1^{N_0-1}$ for which $P_{MM_0N_0}(Z_1^{N_0-1}, M_0 N_0) < \frac{1}{M_0 N_0}$ (hence yielding $K_1(Z_1^{M_0 N_0}, K) < N_0 - 1$, where $K = M_0 N_0$). Note that for any vector $Z_1^{M_0 N_0}$, the parameter t that determines $K_1(Z_1^{M_0 N_0}, K)$, satisfies $t > N_0^2 > N_0$.

Now, $|P_{MM_0N_0}(Z_1^{N_0}, M_0 N_0) - P_{MM_0N_0}(Z_1^{N_0}, N_0)| \leq \frac{M_0 N_0}{M M_0 N_0} \leq \frac{1}{{N_0}^2} = D(N_0)$. By Lemma 2.7 in [5, page 33], for any two probability distributions $P(Z_1^{N_0})$ and $Q(Z_1^{N_0})$,

$$\left| - \sum_{Z_1^{N_0} \in \mathbf{A}^{N_0}} P(Z_1^{N_0}) \log \frac{P(Z_1^{N_0})}{Q(Z_1^{N_0})} \right| \leq d \log \left(\frac{\mathbf{A}^{N_0}}{d} \right)$$

where $d = \max_{Z_1^{N_0} \in A^{N_0}} |P(Z_1^{N_0}) - Q(Z_1^{N_0})|$. Hence,

$$\left| H_{MM_0N_0}(N_0, M_0 N_0) - H_{MM_0N_0}(N_0, N_0) \right| \leq \frac{1}{{N_0}^2} \left[N_0 \log A + 2 \log N_0 \right]$$

and therefore,

$$H_{MM_0N_0}(N_0, N_0) \geq H_u(\mathbf{X}, N_0) - 2\epsilon - \frac{1}{{N_0}^2}\Big[N_0 \log A + 2\log N_0\Big] - \frac{\log A}{N_0}$$

which, by Eqs. (6), (7) and (8) and by setting $\epsilon = \frac{1}{N_0}$, proves Lemma 1.

Proof of Theorem 1: Consider the following construction of X_1^{NM}: Let h be an arbitrary small positive number and ℓ be a positive integer, where δ and ℓ satisfy $N = \ell 2^{h\ell}$, and assume that ℓ divides N.

1) Let $S_{\ell,h}$ be a set of some $T' = \frac{N}{\ell} = 2^{h\ell}$ distinct ℓ-vectors from $\mathbf{A}^\ell$.

2) Generate a concatenation Z_1^N of the T' distinct ℓ-vector in $S_{\ell,h}$.

3) Return to step 2 for the generation of the next N-block.

Now, by construction, for M consecutive N-blocks, $\rho_L(\mathbf{X}, N, M) \geq \frac{1}{M}$ and

$$\frac{1}{N} \geq P_{j,MN}(Z_1^\ell, N) \geq \frac{M-1+N}{MN}; j = 1, 2, ...N.$$

Thus, by construction, $P_{MN}(Z_1^\ell, N) \geq \frac{M-1+N}{MN}$. Furthermore, there exists a positive integer $N_0 = N_0(h)$ such that for any $N \geq N_0$,

$$H_{MN}(\ell, N) \leq \frac{\log N}{\ell} \leq 2h$$

where

$$H_{MN}(\ell, N) = -\frac{1}{\ell} \sum_{Z_1^N \in \mathbf{A}^N} P_{MN}(Z_1^\ell, N) \log P_{MN}(Z_1^\ell, N)\,.$$

Observe that any vector $Z_{j-i}^j; i+1 \leq j \leq MN; 1 \leq i \leq \ell - 1$, except for a subset of instances j with a total empirical probability measure of at most $\frac{1}{2^{h\ell}}$, is therefore a suffix of $Z^j_{j-K_1(X_1^N, \hat{K})}$ where $\hat{K} = N\frac{M}{M-1}$ and that $K_1(X_1^N, \hat{K}) \leq t$ for any $N > N_0(h)$, where $t = \lceil (\log N)^2 \rceil$. Thus, by applying the chain-rule to $H_{MN}(\ell, N)$, by the convexity of the entropy function and by Eq. (5),

$$H_u(N, \hat{K}, M) \leq H_{MN}(Z_1|Z^0_{-\ell+1}, N)_{\leq} H_{MN}(\ell, N) \leq 2h \tag{10}$$

Also, $\limsup_{M\to\infty} H_u(N, \hat{K}, M) = \limsup_{M\to\infty} H_u(\hat{K}, \hat{K}, M-1)$.

Consider now the class $\sigma_{\ell,h}$ of all sets like $S_{\ell,h}$ that consists of $2^{h\ell}$ distinct ℓ-vectors. The next step is to establish that no compression for N-sequences which consist of the $2^{h\ell}$ distinct ℓ vectors that are selected from some member in the class $\sigma_{\ell,h}$ is possible, at least for some such N-sequences.

Let the *normalized* length-function $\bar{L}(Z_1^N)$ be defined by:

$$\hat{L}(Z_1^N) = -\log \frac{2^{-L(Z_1^N)}}{\sum_{Z_1^N \in \mathbf{A}^N} 2^{-L(Z_1^N)}} .$$

Clearly, $\bar{L}(Z_1^N) \leq L(Z_1^N)$ since $L(Z_1^N)$ satisfies the Kraft inequality while $\bar{L}(Z_1^N$ satisfies it with equality, since $2^{-\bar{L}(Z_1^N)}$ is a probability measure. Then,

$$L(Z_1^N) \geq \bar{L}(Z_1^N) = \sum_{i=0}^{\frac{N}{\ell}-1} \bar{L}(Z_{i\ell+1}^{(i+1)\ell} | Z_1^{i\ell})$$

where

$$\bar{L}(Z_{i\ell+1}^{(i+1)\ell} | Z_1^{i\ell}) = \bar{L}(Z_1^{(i+1)\ell}) - \bar{L}(Z_1^{i\ell})$$

is a (normalized) conditional length-function that, given $X_1^{i\ell}$, satisfy the Kraft inequality with equality, since $2^{-\bar{L}(Z_{i\ell+1}^{(i+1)\ell} | Z_1^{i\ell})}$ is a conditional probability measure.

Lemma 2 *For any $h > 0$, any $N \geq N_0 = N_0(\ell, h)$ and any $\bar{L}(X_1^\ell | X_{-N+1}^0)$ there exists a set of $2^{h\ell}$ ℓ-vectors such that*

$$\sum_{X_1^\ell \in \mathbf{A}^\ell} P_{1,N}(X_1^\ell, \ell) \bar{L}(X_1^\ell | X_{-N+1}^0) \geq \ell(1-\delta)(\log A - \delta)$$

for all X_{-N+1}^0 which are concatenations of ℓ-vectors from $S_{\ell,h}$ as described above.

Proof of Lemma 2: The number of possible sets $S_{\ell,h}$ that may be selected from the A^ℓ ℓ vectors over $\mathbf{A}$ is:

$$M_{\ell,h} = \binom{2^{(\log A)\ell}}{2^{h\ell}}$$

Given a particular $\bar{L}(X_1^\ell | X_{-N+1}^0)$, consider the collection $\mathbf{M}_{\ell,h,\delta | X_{-N+1}^0}$ of all sets $S_{\ell,h,\delta}$ that consist of at most $(1-\delta)2^{h\ell}$ vectors selected from the set of vectors $\hat{X}_1^\ell$ for which $L(\hat{X}_1^\ell | X_{-N+1}^0) \geq (\log A - \delta)\ell$ (observe that there are at least $2^{\log A\ell} - 2^{(\log A - \delta)\ell}$ such vectors).

The collection $\mathbf{M}_{\ell,h,\delta | X_{-N+1}}$ is referred to as the collection of "good" sets $S_{\ell,h,\delta}$ (i.e. sets yielding $\hat{L}(X_1^\ell | X_{-N+1}^0) \leq (\log A - \delta)\ell$).

It will now be demonstrated that $\left|\sum_{X^0_{-N+1}\in\mathbf{A}^N}\mathbf{M}_{\ell,h,\delta|X^0_{-N}}\right|$ is exponentially smaller than $M_{\ell,h}$ if $N<\delta 2^{h\ell}(1-h)\ell$. Hence, for any conditional length- function $L(X_1^\ell|X^0_{-N+1})$ and any $X^0_{-N+1})\in\mathbf{A}^N$, most of the sets $S_{\ell,h}\in\mathbf{M}_{\ell,h}$ will not contain a "good" $S_{\ell,h,\delta}\in\mathbf{M}_{\ell,h,\delta|X^0_{-N+1}}$ and therefore less than $\delta 2^{h\ell}$ ℓ-vectors out of the $2^{h\ell}$ ℓ-vectors in $S_{\ell,h}$ will be associated with an $L(X_1^\ell|X^0_{-N+1})<(\log A-\delta)\ell$.

The cardinality of $\mathbf{M}_{\ell,h,\delta|(X^0_{-N+1}}$ is upper bounded by: $\sum_{j=\delta 2^{h\ell}}^{2^{h\ell}}\binom{2^{\ell\log A}-2^{(\log A-\delta)\ell}}{(2^{h\ell}-j)}\binom{2^{(\log A-\delta)\ell}}{j}$. Now, by [3], one has for a large enough positive integer n,

$$\log_2\binom{n}{pn}=[h(p)+\epsilon(n)]n$$

where $h(p)=-p\log_2 p-(1-p)\log_2(1-p)$ and where $\lim_{n\to\infty}\epsilon(n)=0$.

Thus,

$$\log\frac{\sum_{X^0_{-N+1}\in\mathbf{A}^N}\mathbf{M}_{\ell,h,\delta|X^0_{-N+1}}}{M_{\ell,h}}\leq-\delta 2^{h\ell}(1-h)\ell+N+\epsilon'(N)N \tag{11}$$

where $\lim_{N\to\infty}\epsilon'(N)=0$. Therefore, if $N<\delta 2^{h\ell}(1-h)\ell$, there exists some $S_{\ell,h}$ for which,

$$\sum_{X_1^\ell\in\mathbf{A}^\ell}2^{-h\ell}\hat{L}(X_1^\ell|X^0_{-N+1})\geq\ell(1-\delta)(\log A-\delta)$$

for all N-vectors $X^0_{-N+1}\in\mathbf{A}^N$.

Hence by construction, there exists some $S_{\ell,h}$ for which,

$$L(Z_1^N)\geq\bar{L}(Z_1^N)=\sum_{i=0}^{\frac{N}{\ell}-1}\bar{L}\Big(Z_{i\ell+1}^{(i+1)\ell}|X_1^{i\ell}\Big)\geq N\Big[(1-\delta)(\log A-\delta)+\epsilon'(N)\Big]$$

This completes the proof of Lemma 2 and setting $h=\delta$, the proof of Theorem 2.

Proof of Theorem 3: It follows from the construction of of the universal compression algorithm that is associated with Theorem 2 above that $N[H_u(Z_1^N,T_u,t)+O(N^{-\epsilon})]$ is a proper length-function. Consider the one-to-one mapping of X_1^N with the following length-function:

1) $L(Z_1^{N)})=2+N[H_u(Z_1^N,T_u,t)+O(N^{-\epsilon})]$ if $H_u(Z_1^N,T_u,t)\leq H_\epsilon(N,X_1^m)$, and if $Z_1^N\in S_0(N,X_1^m)=S(N,\epsilon,X_1^m)$
2) $L(Z_1^N)=2+N[H_0(N,X_1^m)]$ if $H_u(Z_1^N,T_u,t)>H_\epsilon(N,X_1^m)$ and if $Z_1^N\in S(N,\epsilon,X_1^m)$
3) Else, $L(Z_1^N)=2+N[H_u(Z_1^N,T_u,t)+O(N^{-\epsilon})]$.

Note that for every $Z_1^N \in S(N, \epsilon, X_1^m)$,

$$L(Z_1^N) \leq 2 + N[H_\epsilon(N, X_1^m)]$$

By by Proposition 1 and by Lemma 1, since $L(Z_1^N)$ is a length-function and by the construction of the universal data-compression algorithm that is associated with Theorem 2, for any $\mathbf{X}$, there exists a positive integer N_0 such that for any $N > N_0$ and *some* $M > M_0 = M_0(N_0)$, $H_u(N, N", M) \leq H_u(\mathbf{X}, N) + \frac{1}{4}(\epsilon)^2$ and,

$$\frac{1}{m-N+1}\sum_{j=0}^{m-N} H_u((X_{j+1}^{j+N}, \bar{T}_u, t), + \frac{2}{N} + O(N^{-\epsilon}) \leq H_u(N, N", M) + \frac{2}{N} + O(N^{-\epsilon})$$

Also, by Proposition 1 and Lemma 1 and for any $N \geq N_0'(\mathbf{X}) \geq N_0$ and any $m \geq MN_0'$,

$$\frac{1}{m-N+1}\sum_{j=0}^{m-N} L(X_{j+1}^{j+N}) \geq H_{MN}(N) \geq H_u(\mathbf{X}, N) - \frac{1}{4}(\epsilon)^2$$

Therefore,

$$\frac{1}{m-N+1}\sum_{j=0}^{m-N} \Delta(X_1^m, X_{j+1}^{j+N}) \leq \frac{1}{2}\epsilon^2 + O(N^{-\epsilon})$$

But, as pointed in the description of the universal classifier above,

$$\Delta(X_1^m, X_{j+1}^{j+N}) + O(N^{-\epsilon}) \geq 0; j = 0, 1, ..., m-N$$

Statement 1) in Theorem 3 then follows by the Markov inequality and by setting $\epsilon' = \frac{1}{2}\epsilon^2 + O(N^{-\epsilon})$.

Next, a training sequence X_1^m is constructed, for which $H_{\epsilon,min}(N, X_1^m) > 0$ and where the $2^{NH_{\epsilon,min}(N,X_1^m)}$ "typical" N-vectors in X_1^m are equiprobable. At the same time, $\hat{H}_\epsilon(N, Z_1^N) \leq H_{\epsilon,min}(N, X_1^m) + (\epsilon)^2$, where ϵ is an arbitrarily small positive number.

Hence, if N' is a positive integer that satisfies $\log N' < \log N$, any classifier with a storage complexity of $O(N')$ can store only an exponentially small fraction of the $2^{NH_{\epsilon,min}(N,X_1^m)}$ typical N-vectors that should be accepted and therefore, for any ϵ-efficient classifier, $\hat{H}_\epsilon(N, Z_1^N) \geq \log A - \epsilon$.

This will lead to statement 2) of Theorem 3.

Let B and ℓ be positive integers satisfying $B > \frac{\ell^{3\epsilon}}{\epsilon}$ and let $R_{\ell,h}$ be the set of all the ℓ-vectors in $S_{\ell,h}$ and their cyclic shifts, where $S_{\ell,h}$ is described in the proof of Theorem 1 above. Consider training sequences X_1^m that are generated by B repetitions of an ℓ-vectors in $S_{\ell,h}$, followed by B repetitions of yet another ℓ vector in $R_{\ell,h}$ and so on, until $R_{\ell,h}$ is exhausted. Thus, $m = (\ell)^2 2^{h\ell} B$.

Let $N = \epsilon \ell B$ and assuming that N divides m, let $M = \frac{m}{N}$. It then follows that $0 < H_{\epsilon,min}(N, X_1^m) \leq (h + \frac{\log \ell}{\ell})$. Also, it follows that $P_{MN}(Z_j^{j+\ell}) \geq \frac{1}{(1-\epsilon)\ell}$ for most instances j, except for a subset of instances that has an empirical probability of at most ϵ. Thus, similar to the proof of Theorem 1 above, it follows that there exists a positive integer ℓ_0 such that for any $\ell \geq \ell_0$, $H_{MN}(\ell, N) \leq h + \frac{\log \ell}{\ell} + \epsilon \log A$ and also, for small enough ϵ, $H_u(N, K, M) \leq H_u(N, \frac{\ell}{1-\epsilon}, M) < 2h$, where $K = N$" and where $N" = N^{1-2\epsilon} = \ell^{(1+3\epsilon)(1-2\epsilon)} \geq \frac{\ell}{1-\epsilon}$. Therefore, $\Delta(X_1^m, X_{j+1}^{j+N}) \leq 2h$.

Hence, by setting $h = \epsilon^2$ and $\epsilon' = \epsilon^2 + O(N^{-\epsilon})$, and by the Markov inequality, $\hat{H}_\epsilon(N, Z_1^N) \leq H_{\epsilon,min}(N, X_1^m) + (\epsilon)^2$. Also, the $2^{NH_{\epsilon,min}(N,X_1^m)}$ "typical" N-vectors in X_1^m are equiprobable, which completes the proof of Theorem 3.

References

[1] J. Ziv, A. Lempel, "Compression of Individual Sequences via Variable-Rate Coding", *IEEE Trans. Inf. Theory*, vol. IT–24, no. 5, pp. 530–536, Sept. 1978.

[2] A. Martin, G. Seroussi and M. Weinberger, "Linear Time Universal Coding and Time Reversal of Tree Sources via FSM Closure", *IEEE Trans. Inf. Theory*, vol. 50, no. 7, July 2004.

[3] R.B. Ash, "Information Theory", Dover, New York, 1965; pp. 113.

[4] F.M.J. Willems, Y.M. Shtarkov, T.J. Tjalkens, "The Context-Tree Weighting Method: basic properties", *IEEE Trans. Inf. Theory*, vol. IT-41, no. 3, pp. 653–664, May 1995.

[5] I. Csiszar, J. Korner, "Information Theory:Coding Theorems for Discrete Memoryless Systems", Academic Press, 1981.

[6] P. C. Shields, "The Interactions Between Ergodic Theory and Information Theory", *IEEE Trans. Inf. Theory*, vol. IT-44, no. 6, 1998.

[7] D. Ron, Y. Singer, N. Tishbi, "The power of Amnesia: Learning Probabilistic Automata with Variable Memory Length", *Machine Learning*, vol. 25, no. 2-3, pp. 117–149, 1996.

[8] A. Apostolico and G. Bejerano, "Optimal Amnesic Probabilistic Automata or how to learn and classify Proteins in Linear Time and Space", *Proc. of ACM Recomb.*, pp. 25-32, 2000.

[9] M. Feder, N. Merhav,M. Gutman Universal, "prediction of individual sequences". *IEEE Trans. Inf. Theory*, vol. IT-38, no. 4, pp. 1258-1270, July 1992.

[10] j. Rissanen, "Fast universal coding with context models" *IEEE Trans. Inf. Theory*, vol. IT-45, no. 4, pp. 1065-1071, May 1999.

[11] J.Ziv "Classification with Finite Memory Revisited" Submitted to the *IEEE Trans. Inf. Theory.*

[12] J. Ziv, "An efficient universal prediction algorithm for unknown sources with limited training data", IEEE Trans. Inf. Theory, vol. IT-48, pp. 1690-1693, June 2002.

and Correction to: "An efficient universal prediction algorithm for unknown sources with limited training data", *IEEE Trans. Inf. Theory*, vol. IT-50, pp. 1851-1852, August 2004.

[13] P. Jacquet, W. Szpankowski, I Apostol, "universal predictor based on pattern matching" *IEEE Trans. Inf. Theory*, Volume 48, Issue 6, June 2002 pp. 1462 - 1472

[14] J.Ziv, N.Merhav, "a measure of relative entropy between individual sequences with applications to universal classification" *IEEE Trans. Inf. Theory*, vol IT-39 1993, pp. 1280-1292

[15] M.Ridley "The search for LUCA." Natural History November 2000, pp. 82-85