

Symbol-Based Modeling and Coding of Block Markov Sources

Dániel A. Nagy András György Tamás Linder

IEEE Transactions on Information Theory

Submitted: September 2, 2005

Abstract

Industry-standard lossless compression algorithms (such as LZW) are usually implemented so that they work on bytes as symbols. Experiments indicate that data for which bytes are not the natural choice of symbols compress poorly using these implementations, while algorithms working on a bit level perform reasonably on byte-based data in addition to having computational advantages resulting from operating on a small alphabet. In this paper, we offer an information-theoretic explanation to these experimental results by assessing the redundancy (which is approximated by the divergence rate of two source distributions) of a bit-based model when applied to a byte-based source. More specifically, we study the problem of approximating a block Markov source with higher order Markov sources, and show that the divergence rate between a block Markov source and the best-matching higher order Markov model for that source converges to zero exponentially fast as the memory of the model increases. This result is applied to obtain bounds on the redundancy of certain symbol-based universal codes when they are used for byte-aligned sources.

Index Terms: lossless coding, block Markov sources, higher order Markov modeling, binary codes, byte-aligned sources.

D. A. Nagy and T. Linder are with the Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada K7L 3N6 (email: {nagydani}{linder}@mast.queensu.ca). A. György is with the Informatics Laboratory, Computer and Automation Research Institute of the Hungarian Academy of Sciences, Lágymányosi út 11, 1111 Budapest, Hungary (email: gya@szit.bme.hu). During part of this work he was also with the Department of Mathematics and Statistics, Queen's University. This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada, the NATO Science Fellowship of Canada, and the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

1 Motivation

The goal of lossless data compression is to represent digital data using as few binary symbols (bits) as possible with a subsequent error-free reconstruction. In many cases, very little prior information is available about the data to be compressed and one is compelled to use universal (adaptive) data compression algorithms. For historical reasons, most digital data are represented as sequences of bytes (eight-bit blocks), but there is a substantial amount of data for which this byte-aligned representation is not natural (e.g., genetic code, where proteins are encoded by sequences of 3 bases, which in turn can be of four kinds, thus one protein is described by 6 bits). Yet, the majority of compression algorithm implementations have the assumption of byte-alignment hard-coded into them, making them surprisingly inefficient for data not aligned to byte boundaries.

Implementing data-compression algorithms on the bit level has several advantages from a computational point of view. Moreover, experimental data suggests that the penalty for not taking byte-alignment into account for many byte-aligned sources seems acceptably low [1]. Specifically, in our experiments a clearly sub-optimal bit-level BWT-based [2] compressor has significantly outperformed industry standard compressors (`gzip` and `bzip2`) on data that were not aligned to byte boundaries, while being only slightly inferior in compressing byte-aligned sources. For example, when applied to data in the Calgary Corpus [3], our compressor consistently produced compressed files that were only around 15% larger than those for `bzip`, a similar compression algorithm implemented on bytes.

Motivated by these practical observations, in present paper we evaluate this penalty in an information-theoretic setting. Specifically, under Markovian assumptions we investigate the excess of encoding rate resulting when a lossless code that is optimized for a source with atomic symbols (e.g., bits) is applied to a source with symbols that are blocks of these atomic symbols (e.g., bytes).

The minimum achievable rate for lossless coding is the entropy rate of the source [4]. The excess of code rate over the entropy rate is called the redundancy of the code. This is the quantity that needs to be minimized when designing a lossless code. If the code is optimal for the model distribution, the relative entropy rate [5] between the model distribution and that of the source approximates the rate redundancy of the code with respect to the source.

In Section 3 we analyze the divergence rate between a block Markov source and the best fitting higher order Markov model. The main result here gives an explicit formula for this divergence rate which implies that higher order Markov models can efficiently model block Markov sources. In Section 4 we show that the convergence to zero of the divergence rate is in fact exponential in the order of the memory of the Markov model. Finally, in Section 5 we upper bound the redundancy on block Markov sources of a large class of codes that are universal for higher-order Markov sources.

This bound makes it possible to choose the order of the Markov model in a way that optimizes a complexity-redundancy tradeoff.

2 Preliminaries

For any sequence of random variables $\{X_n\}_{n=0}^\infty = X_0, X_1, \dots, X_n, \dots$ and for any $i \geq j$, the segment $(X_i, X_{i+1}, \dots, X_j)$ will be denoted by X_i^j . We allow j to be infinite; for example, we write X_0^∞ for the entire sequence $\{X_n\}_{n=0}^\infty$. A similar convention applies to deterministic sequences which are usually denoted using lower case letters.

For any pair of discrete random variables Z and V taking values in the finite sets \mathcal{Z} and \mathcal{V} , respectively, let $P_Z(z) = \Pr(Z = z)$ and $P_{Z|V}(z|v) = \Pr(Z = z|V = v)$ for all $z \in \mathcal{Z}$ and $v \in \mathcal{V}$. If $\mathcal{Z} = \mathcal{V}$, the relative entropy (Kullback-Leibler divergence) between Z and V is defined as

$$D(Z\|V) = D(P_Z\|P_V) = \sum_{z \in \mathcal{Z}} P_Z(z) \log \frac{P_Z(z)}{P_V(z)}$$

where \log denotes base 2 logarithm. $D(P_Z\|P_V)$ is nonnegative and equals zero if and only if $P_Z = P_V$ [5]. For sequences of random variables Z_0^∞ and V_0^∞ , the divergence rate is defined as

$$D(Z_0^\infty\|V_0^\infty) = \lim_{n \rightarrow \infty} \frac{1}{n} D(Z_0^{n-1}\|V_0^{n-1})$$

provided the limit exists.

The sequence of random variables X_0^∞ taking values in the finite alphabet \mathcal{A} is called a block- N Markov source if for every nonnegative integer i and block of symbols $x_0^{(i+1)N-1} \in \mathcal{A}^{(i+1)N}$,

$$\begin{aligned} P_{X_{iN}^{(i+1)N-1}|X_0^{iN-1}} \left(x_{iN}^{(i+1)N-1} | x_0^{iN-1} \right) \\ &= P_{X_{iN}^{(i+1)N-1}|X_{(i-1)N}^{iN-1}} \left(x_{iN}^{(i+1)N-1} | x_{(i-1)N}^{iN-1} \right) \\ &= P_{X_N^{2N-1}|X_0^{N-1}} \left(x_{iN}^{(i+1)N-1} | x_{(i-1)N}^{iN-1} \right). \end{aligned}$$

The sequence of random variables Y_0^∞ taking values in \mathcal{A} is called an m th order Markov source if for every nonnegative integer i and $x_0^{m+i} \in \mathcal{A}^{i+m+1}$,

$$P_{Y_{i+m}|Y_0^{i+m-1}}(x_{i+m} | x_0^{i+m-1}) = P_{Y_m|Y_0^{m-1}}(x_{i+m} | x_i^{m-1}).$$

A binary block code of length n for the source alphabet \mathcal{A} is given by a function $f_n : \mathcal{A}^n \rightarrow \{0, 1\}^*$, which maps any source vector $x \in \mathcal{A}^n$ to the binary string $f(x)$. The length function $\ell_n : \mathcal{A}^n \rightarrow \mathbb{N}$ associated with f_n gives for each x the length of the corresponding binary string, that is, $\ell_n(x) = |f_n(x)|$. We require f_n to be uniquely decodable, that is, for $x_1, \dots, x_j, y_1, \dots, y_k \in \mathcal{A}^n$,

$f_n(x_1)f_n(x_2)\dots f_n(x_j) = f_n(y_1)f_n(y_2)\dots f_n(y_k)$ if and only if $j = k$ and $x_i = y_i, i = 1, \dots, j$, where for two binary strings s_1 and s_2 , s_1s_2 denotes their concatenation. It is well known [5] that if f_n is uniquely decodable then its length function ℓ satisfies the Kraft inequality

$$\sum_{x \in \mathcal{A}^n} 2^{-\ell_n(x)} \leq 1.$$

Moreover, for any such code there exists a prefix code with the same length function, and also there exists another prefix code f'_n with length function ℓ'_n such that $\ell'_n(x) \leq \ell_n(x)$ for all $x \in \mathcal{A}^n$, and the equality holds for ℓ'_n in the Kraft inequality, that is, $\sum_{x \in \mathcal{A}^n} 2^{-\ell'_n(x)} = 1$. Therefore, without loss of generality, in the rest of the paper we consider only codes for which the Kraft inequality holds with equality. Therefore, the *coding distribution* of f_n , defined as

$$P_{f_n}(x) = 2^{-\ell_n(x)}$$

for each $x \in \mathcal{A}^n$, is a proper probability distribution.

The redundancy of the code f_n with length function ℓ_n for the random vector X_0^{n-1} is defined as

$$R_n = E\ell_n(X_0^{n-1}) - H(X_0^{n-1}) = E\left(\ell_n(X_0^{n-1}) - \log P_{X_0^{n-1}}(X_0^{n-1})\right)$$

the difference of the expected code length $E\ell_n(X_0^{n-1})$ and the entropy

$$H(X_0^{n-1}) = - \sum_{x \in \mathcal{A}^n} P_{X_0^{n-1}}(x) \log P_{X_0^{n-1}}(x).$$

Note that $R_n \geq 0$, and if Y_0^{n-1} is distributed according to P_{f_n} , then

$$R_n = D(X_0^{n-1} \| Y_0^{n-1}).$$

Similarly, for any distribution P_n over \mathcal{A}^n , one can construct a prefix code with length function $\ell'_n(x) = -\lceil \log P_n(x) \rceil$. The redundancy of this code can be bounded as

$$R'_n = E\ell'_n(X_0^{n-1}) - H(X_0^{n-1}) \leq D(X_0^{n-1} \| \widehat{Y}_0^{n-1}) + 1$$

where \widehat{Y}_0^{n-1} is distributed according to P_n .

A binary source code for an infinite source X_0^∞ taking values in the alphabet \mathcal{A} is given by a sequence of block- n codes f_n . Without loss of generality we assume that for each f_n equality holds in the Kraft inequality. If the coding distributions P_{f_n} are compatible in the sense that there is an \mathcal{A} -valued random process Y_0^∞ such that the distribution of Y_0^{n-1} is P_{f_n} for all n , then the redundancy rate of the code is given as

$$\lim_{n \rightarrow \infty} \frac{1}{n} R_n = \lim_{n \rightarrow \infty} \frac{1}{n} D(X_0^{n-1} \| Y_0^{n-1}) = D(X_0^\infty \| Y_0^\infty)$$

provided the limit exists [5],[6]. If X_0^∞ is a block- N stationary block- N Markov source and Y_0^∞ is a stationary m th order Markov source, then both sources are block stationary block- mN Markov sources; for such sources the limit always exists [7].

In the sequel we will alternately use the *code* for either a block- n code f_n , or a sequence of codes $\{f_n\}_{n=1}^\infty$.

3 Approximation of block Markov sources

In this section we want to find the best m th order Markovian approximation of a block- N stationary block- N Markov source X_0^∞ in the sense that we look for an m th order Markov source Y_0^∞ achieving the minimum

$$\bar{D}_m \triangleq \min\{\bar{D}(X_0^\infty \| Y_0^\infty) : Y_0^\infty \text{ is } m\text{th order Markov}\}.$$

Clearly, without loss of generality we may assume that Y_0^∞ is stationary.

Let $\{X_n\}_{n=-\infty}^\infty$ be the two-sided block- N stationary extension of $\{X_n\}_{n=0}^\infty$, and let $\{Y_n\}_{n=-\infty}^\infty$ be the two-sided stationary extension of $\{Y_n\}_{n=0}^\infty$. The minimizing $\{Y_n\}$ and the minimum divergence rate will be expressed in terms of the random variables

$$U_j = X_{j-m+\tau}, \quad j = 0, 1, 2, \dots$$

where τ is a random variable that is uniformly distributed on $\{0, 1, \dots, N-1\}$ and is independent of $\{X_n\}$. Notice that $\{U_j\}$ can be seen as a stationary version of the (only) block- N stationary source $\{X_n\}$. With this in mind, it is intuitively clear that the best m th order Markovian approximation of $\{U_n\}$, which has the same m th order conditional distributions as $\{U_n\}$, will also be the best approximation for $\{X_n\}$. This statement is formalized in the next theorem.

Theorem 1 *Given a block- N Markov source X_0^∞ , the relative entropy rate $\bar{D}(X_0^\infty \| Y_0^\infty)$ is minimized over all stationary m th order Markov sources Y_0^∞ if and only if $P_{Y_m|Y_0^{m-1}} = P_{U_m|U_0^{m-1}}$. The minimum relative entropy rate is given for all $m \geq 2N$ by*

$$\bar{D}_m = I(\tau; U_m | U_0^{m-1})$$

the conditional mutual information between τ and U_m given U_0^{m-1} . Moreover, there is a stationary version \hat{Y}_0^∞ of Y_0^∞ such that $P_{\hat{Y}_0^m} = P_{U_0^m}$.

Expressing conditional mutual information in terms conditional entropies as

$$I(\tau; U_m | U_0^{m-1}) = H(\tau | U_0^{m-1}) - H(\tau | U_0^m)$$

we obtain

$$\begin{aligned}
& \sum_{m=2N}^{\infty} I(\tau; U_m | U_0^{m-1}) \\
&= \sum_{m=2N}^{\infty} \left(H(\tau | X_{\tau-m}^{\tau-1}) - H(\tau | X_{\tau-m}^{\tau}) \right) \\
&\leq H(\tau | X_{\tau-2N}^{\tau-1}) - \liminf_{m \rightarrow \infty} H(\tau | X_{\tau}^{\tau-m}) \leq \log N
\end{aligned}$$

where the first inequality follows since we clearly have $H(\tau | X_{\tau-m-1}^{\tau-1}) = H(\tau | X_{\tau-m}^{\tau})$.

Thus we obtain the following corollary which states that the block Markov source can be arbitrarily closely approximated by higher-order Markov models by increasing the model order.

Corollary 1 *The minimum relative entropy rate \bar{D}_m satisfies*

$$\sum_{m=2N}^{\infty} \bar{D}_m \leq \log N.$$

In particular

$$\lim_{m \rightarrow \infty} \bar{D}_m = 0.$$

Remark The fact that \bar{D}_m converges to zero as $m \rightarrow \infty$ is not very surprising in view of the fact that the divergence rate between a stationary process and its best m th order Markov approximation asymptotically vanishes as $m \rightarrow \infty$ (see, e.g., [7]). Note, however, that X_0^∞ is non-stationary, and that the theorem gives an explicit expression for the optimum approximating process and a characterization of the resulting minimum divergence rate \bar{D}_m . In the next section we will use this result to determine the rate at which \bar{D}_m converges to zero.

Proof of Theorem 1 For all $n > m$ we have from the chain rule for the relative entropy [5]

$$\begin{aligned}
& D(P_{X_0^n} \| P_{Y_0^n}) \\
&= \sum_{i=m}^n D(P_{X_i | X_0^{i-1}} \| P_{Y_i | Y_0^{i-1}}) + D(P_{X_0^{m-1}} \| P_{Y_0^{m-1}})
\end{aligned}$$

where

$$D(P_{X_i | X_0^{i-1}} \| P_{Y_i | Y_0^{i-1}}) = \sum_{a_0^i \in \mathcal{A}^{i+1}} P_{X_0^i}(x_0^i) \log \frac{P_{X_i | X_0^{i-1}}(a_i | a_0^{i-1})}{P_{Y_i | Y_0^{i-1}}(a_i | a_0^{i-1})}.$$

Observe that if $m \geq 2N$, then for any $i \geq m$,

$$P_{X_i | X_0^{i-1}}(\cdot | x_0^{i-1}) = P_{X_i | X_{i-m}^{i-1}}(\cdot | x_{i-m}^{i-1})$$

and

$$P_{Y_i|Y_0^{i-1}}(\cdot|y_0^{i-1}) = P_{Y_m|Y_0^{m-1}}(\cdot|y_0^{m-1}).$$

Therefore

$$\begin{aligned} D(P_{X_i|X_0^{i-1}} \| P_{Y_i|Y_0^{i-1}}) &= \sum_{a \in \mathcal{A}^i} P_{X_0^{i-1}}(a) D\left(P_{X_i|X_0^{i-1}}(\cdot|a) \| P_{Y_i|Y_0^{i-1}}(\cdot|a)\right) \\ &= \sum_{b \in \mathcal{A}^m} P_{X_{i-m}^{i-1}}(b) D\left(P_{X_i|X_{i-m}^{i-1}}(\cdot|b) \| P_{Y_m|Y_0^{m-1}}(\cdot|b)\right) \\ &= \sum_{b \in \mathcal{A}^m} P_{X_{t-m}^{t-1}}(b) D\left(P_{X_t|X_{t-m}^{t-1}}(\cdot|b) \| P_{Y_m|Y_0^{m-1}}(\cdot|b)\right) \end{aligned}$$

where $t = i \bmod N$. Denoting the last sum by S_t , we obtain

$$\lim_{n \rightarrow \infty} \frac{1}{n+1} D(P_{X_0^n} \| P_{Y_0^n}) = \lim_{n \rightarrow \infty} \frac{1}{n+1} \sum_{i=m}^n D(P_{X_i|X_0^{i-1}} \| P_{Y_i|Y_0^{i-1}}) = \frac{1}{N} \sum_{t=0}^{N-1} S_t.$$

Let τ denote a uniform random variable over $\{0, 1, \dots, N-1\}$ that is independent of the pair $(\{X_n\}, \{Y_n\})$ pair, and define the random vectors $U_0^m = X_{\tau-m}^\tau$ and $V_0^m = Y_{\tau-m}^\tau$. Then we can rewrite the relative entropy rate as

$$\begin{aligned} \bar{D}(X_0^\infty \| Y_0^\infty) &= \sum_{t=0}^{N-1} P_\tau(t) \sum_{b \in \mathcal{A}^m} P_{U_0^{m-1}|\tau}(b|t) \\ &\quad \cdot D\left(P_{U_m|U_0^{m-1},\tau}(\cdot|b,t) \| P_{V_m|V_0^{m-1},\tau}(\cdot|b,t)\right) \\ &= \sum_{t=0}^{N-1} P_\tau(t) \sum_{b \in \mathcal{A}^m} P_{U_0^{m-1}|\tau}(b|t) \\ &\quad \cdot \sum_{x \in \mathcal{A}} P_{U_m|U_0^{m-1},\tau}(x|b,t) \log \frac{P_{U_m|U_0^{m-1},\tau}(x|b,t)}{P_{Y_m|Y_0^{m-1}}(x|b)} \\ &= \sum_{t=0}^{N-1} \sum_{b \in \mathcal{A}^m} \sum_{x \in \mathcal{A}} P_{U_0^m,\tau}(b,x,t) \\ &\quad \cdot \log \frac{P_{\tau|U_0^m}(t|b,x) P_{U_m|U_0^{m-1}}(x|b)}{P_{Y_m|Y_0^{m-1}}(x|b) P_{\tau|U_0^{m-1}}(t|b)} \\ &= \sum_{t=0}^{N-1} \sum_{b \in \mathcal{A}^m} \sum_{x \in \mathcal{A}} P_{U_0^m,\tau}(b,x,t) \log \frac{P_{\tau|U_0^m}(t|b,x)}{P_{\tau|U_0^{m-1}}(t|b)} \\ &\quad + \sum_{t=0}^{N-1} \sum_{b \in \mathcal{A}^m} \sum_{x \in \mathcal{A}} P_{U_0^m,\tau}(b,x,t) \log \frac{P_{U_m|U_0^{m-1}}(x|b)}{P_{Y_m|Y_0^{m-1}}(x|b)}. \end{aligned}$$

Observe that only the second term of the last expression depends on the choice of $\{Y_n\}$. Since this term is equal to $D(P_{U_m|U_0^{m-1}}\|P_{Y_m|Y_0^{m-1}})$ (so it is nonnegative), it is uniquely minimized by the choice $P_{Y_m|Y_0^{m-1}} = P_{U_m|U_0^{m-1}}$. With this optimum choice the second term vanishes, so

$$\begin{aligned}
\bar{D}_m &= \sum_{t=0}^{N-1} \sum_{b \in \mathcal{A}^m} \sum_{x \in \mathcal{A}} P_{U_0^m, \tau}(b, x, t) \log \frac{P_{\tau|U_0^m}(t | b, x)}{P_{\tau|U_0^{m-1}}(t | b)} \\
&= \sum_{t=0}^{N-1} \sum_{b \in \mathcal{A}^m} \sum_{x \in \mathcal{A}} P_{U_0^m, \tau}(b, x, t) \log P_{\tau|U_0^m}(t | b, x) \\
&\quad - \sum_{t=0}^{N-1} \sum_{b \in \mathcal{A}^m} P_{U_0^{m-1}, \tau}(b, t) \log P_{\tau|U_0^{m-1}}(t | b) \\
&= H(\tau | U_0^{m-1}) - H(\tau | U_0^m) = I(\tau; U_m | U_0^{m-1})
\end{aligned}$$

which was to be shown.

Finally, as $P_{Y_m|Y_0^{m-1}} = P_{U_m|U_0^{m-1}}$ and U_0^∞ is stationary, starting the m th order Markov chain Y_0^∞ from the distribution $P_{U_0^{m-1}}$ results in a stationary version of Y_0^∞ . This proves the last statement of the theorem. \square

From a coding point of view, Theorem 1 states that if a coding procedure is optimal for Y_0^∞ (in the sense that its length functions correspond to the marginal distributions of Y_0^∞), then it can compress X_0^∞ with rate not exceeding the source entropy rate $\bar{H}(X_0^\infty) = \lim_{n \rightarrow \infty} H(X_0^{n-1})$ by more than \bar{D}_m . However, in practical situations such codes are not available, as the distribution of Y_0^∞ is usually not known. Moreover, as the triangle inequality does not hold for divergences, a code which is almost optimal for Y_0^∞ need not be good at all for X_0^∞ . Still, it is reasonable to expect that codes that are universal for the class of m th order Markov sources (that is, perform asymptotically optimally for all sources in the class, including Y_0^∞) will perform well on X_0^∞ . This will be shown (together with convergence rates) in Section 5.

4 Rate of convergence

In this section we examine the rate of convergence at which the minimum relative entropy rate \bar{D}_m converges to 0 in Corollary 1. In fact, we will show that \bar{D}_m vanishes exponentially fast, that is, a block Markov source can be very well approximated by high order Markov sources.

From Theorem 1 we can see that in order to establish that rate of convergence, it is sufficient to estimate the conditional entropy $H(\tau|U_0^m)$. Using Fano's inequality (see, e.g., [5]) we will trace back our problem to the problem of classification of Markov sources. In this latter problem, given finitely many Markov sources, one has to decide which one of them has generated an observed sequence. In previous works it was shown that, under various conditions, this problem can be solved

with exponentially decaying error probability as the length of the observed sequence increases, see, e.g., [8, 9, 10]. However, the conditions in these works are not immediately applicable to our setup. Therefore, first we revisit some results from Csiszár *et al.* [11] concerning large deviations of Markov chains. Based on these results, in Lemma 2 we derive an upper bound on the classification error (using a similar method as in [8]).

Let \mathcal{A} be a finite set, and let Λ denote the set of distributions over \mathcal{A}^2 . The second order type of a sequence $u_0^n = (u_0, \dots, u_n) \in \mathcal{A}^{n+1}$ is the empirical distribution of the pairs (u_k, u_{k+1}) over \mathcal{A}^2 defined by the relative frequencies

$$P_{u_0^n}^{(2)}(v_1, v_2) = \frac{1}{n} |\{k \in \{0, \dots, n-1\} : u_k = v_1, u_{k+1} = v_2\}|, v_1, v_2 \in \mathcal{A}.$$

For any distribution $P(u, v)$ over \mathcal{A}^2 , let $\bar{P}(u) = \sum_{v \in \mathcal{A}} P(u, v)$ denote the marginal distribution of the first coordinate, and for $\bar{P}(u) > 0$, let $P(v|u) = P(u, v)/\bar{P}(u)$. For any stochastic matrix $\{W(v|u)\}_{u, v \in \mathcal{A}}$ (that is, $W(v|u) \geq 0$ for all $u, v \in \mathcal{A}$, and $\sum_{v \in \mathcal{A}} W(v|u) = 1$), let

$$D(P\|W) = \sum_{u, v} P(u, v) \log \frac{P(u, v)}{\bar{P}(u)W(v|u)} = \sum_{u, v} P(u, v) \log \frac{P(v|u)}{W(v|u)}$$

denote the relative entropy between the distributions $P(u, v)$ and $\bar{P}(u)W(v|u)$. For any set of distributions $\Pi \subset \Lambda$ let $\text{cl}\Pi$ denote its closure in Λ under point-wise convergence. Finally, let $\Lambda_0 = \{P \in \Lambda : \sum_{v \in \mathcal{A}} P(u, v) = \sum_{v \in \mathcal{A}} P(v, u), u \in \mathcal{A}\}$ denote the distributions in Λ with equal marginals, and for any stochastic matrix W , let $S(W) = \{P \in \Lambda : P(u, v) = 0 \text{ if } W(v|u) = 0, u, v \in \mathcal{A}\}$. The following lemma is proved in [11].

Lemma 1 ([11, Lemma 2a]) *Assume that $\{X_i\}$ is a Markov chain with finite alphabet \mathcal{A} and transition matrix W , and let $\Pi \subset S(W)$ be arbitrary. Then for every $u \in \mathcal{A}$,*

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log \Pr(P_{X_0^n}^{(2)} \in \Pi | X_0 = u) \leq - \min_{P \in \Lambda_0 \cap \text{cl}\Pi} D(P\|W).$$

Note that the minimum on the right hand side of the above inequality is attainable, as it is the minimum of a continuous function over a compact set.

Following the approach of Natarajan [8], this result easily leads to classification of Markov sources. Assume that the sample X_1, X_2, \dots, X_n is generated with equal probability by one of K stationary Markov sources with transition matrices W_1, \dots, W_K , respectively. The problem is to determine which source has generated the sample. The next lemma provides a classification method for irreducible Markov-chains with exponentially decaying error probability as the sample size grows. (A Markov-chain with transition matrix W is called irreducible if for every pair $(u, v) \in \mathcal{A}^2$ there is a positive integer n such that the element in the (u, v) position of W^n is positive.)

Lemma 2 Let $\{X_{i,n}\}_{n=0}^\infty$, $i = 1, \dots, K$, $K \geq 2$, be independent Markov sources with irreducible transition matrices W_i such that $W_i \neq W_j$ for $i \neq j$. Assume that t is distributed over $\{1, \dots, K\}$ such that $\Pr(t = i) > 0$ for all $i = 1, \dots, K$, and t is independent of the $\{X_{i,n}\}$'s. Finally, assume that we observe the t th Markov source, that is, let $X_n = X_{t,n}$ for $n = 0, 1, \dots$. Define

$$R_i = \{P \in \Lambda : D(P\|W_i) < D(P\|W_j) \text{ for all } j \neq i\},$$

and let $\hat{t}_n = i$ if $P_{X_0^n}^{(2)} \in R_i$ for some $i \in \{1, \dots, K\}$ and let \hat{t}_n be arbitrary otherwise. Then for any $u \in \mathcal{A}$,

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log \Pr(t \neq \hat{t}_n | X_0 = u) \leq - \min_{1 \leq i \leq K} \min_{P \in \Lambda_0 \cap \bar{R}_i} D(P\|W_i) < 0$$

where $\bar{R}_i = \Lambda \setminus R_i$ denotes the complement of R_i .

Proof. It is easy to see that

$$\begin{aligned} \Pr(t \neq \hat{t}_n | X_0 = u) &= \sum_{i=1}^K \Pr(t = i | X_0 = u) \Pr(\hat{t}_n \neq i | t = i, X_0 = u) \\ &= \sum_{i=1}^K \Pr(t = i | X_0 = u) \Pr(\hat{t}_n \neq i | t = i, X_{i,0} = u) \\ &\leq \sum_{i=1}^K \Pr(t = i | X_0 = u) \Pr(P_{X_{i,0}^n}^{(2)} \in \bar{R}_i | X_{i,0} = u) \\ &\leq \max_i \Pr(P_{X_{i,0}^n}^{(2)} \in \bar{R}_i | X_{i,0} = u) \\ &= \max_i \Pr(P_{X_{i,0}^n}^{(2)} \in \bar{R}_i \cap S(W_i) | X_{i,0} = u). \end{aligned}$$

Here the last equality holds as a sequence $X_{i,0}^n$ has zero probability if it contains a transition of probability zero. Now from Lemma 1 it follows that

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log \Pr(P_{X_{i,0}^n}^{(2)} \in \bar{R}_i \cap S(W_i) | X_{i,0} = u) \leq - \min_{P \in \Lambda_0 \cap \text{cl}(\bar{R}_i \cap S(W_i))} D(P\|W_i).$$

Moreover, since $D(P\|W_i) = \infty$ for any $P \notin S(W_i)$, $\bar{R}_i \cap S(W_i)$ is clearly nonempty, and \bar{R}_i is closed, we have

$$\min_{P \in \Lambda_0 \cap \text{cl}(\bar{R}_i \cap S(W_i))} D(P\|W_i) = \min_{P \in \Lambda_0 \cap \bar{R}_i} D(P\|W_i).$$

This yields

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log \Pr(t \neq \hat{t}_n) \leq - \min_i \min_{P \in \Lambda_0 \cap \bar{R}_i} D(P\|W_i).$$

Finally, it remains to show that the right hand side in the above equation is nonzero. Assuming the contrary, we have that for a given i , P_i minimizes $D(P\|W_i)$ in $\Lambda_0 \cap \bar{R}_i$, and $D(P_i\|W_i) = 0$. The latter divergence can be rewritten as

$$D(P_i\|W_i) = \sum_{u:\bar{P}_i(u)>0} \bar{P}_i(u) \sum_v P_i(v|u) \log \frac{P_i(v|u)}{W_i(v|u)} = \sum_{u:\bar{P}_i(u)>0} \bar{P}_i(u) D(P_i(\cdot|u)\|W_i(\cdot|u))$$

where $D(P_i(v|u)\|W_i(v|u))$ denotes the divergence between the two conditional distributions for a given u . Now since $D(P_i\|W_i) = 0$, we have $D(P_i(\cdot|u)\|W_i(\cdot|u)) = 0$ and hence $P_i(v|u) = W_i(v|u)$ for all u with $\bar{P}_i(u) > 0$. Moreover, the irreducibility of W_i implies that if there is a u with $\bar{P}_i(u) = 0$, then there is a u' such that $\bar{P}_i(u') > 0$ and $P_i(u|u') > 0$. Thus, $P_i(u', u) > 0$. Since $P_i \in \Lambda_0$, its marginals are equal, and so $\bar{P}_i(u) = \sum_{\hat{u}} P_i(\hat{u}, u) > 0$, a contradiction. Therefore, $\bar{P}_i(u) > 0$ for all $u \in \mathcal{A}$, and so $P_i(v|u) = W_i(v|u)$ for all $u, v \in \mathcal{A}$. Moreover, as $P_i \notin R_i$, there is some $j \neq i$ such that $D(P_i\|W_j) \leq D(P_i\|W_i) = 0$. Thus $D(P_i\|W_j) = 0$ by the non-negativity of the relative entropy. Then, since W_j is also irreducible, $P_i(v|u) = W_j(v|u)$ for all $u, v \in \mathcal{A}$. Therefore, $W_i = W_j$, a contradiction. \square

Remark. Note that the end of the proof heavily depends on the fact that the Markov chains are irreducible. Indeed, it is easy to construct reducible Markov chains such that it is impossible to distinguish between them with vanishing error probability no matter how large the sample size is. For example, consider the following two transition matrices:

$$W_1 = \begin{pmatrix} .5 & .5 & 0 & 0 \\ .5 & .5 & 0 & 0 \\ 0 & 0 & .5 & .5 \\ 0 & 0 & .5 & .5 \end{pmatrix} \quad \text{and} \quad W_2 = \begin{pmatrix} .5 & .5 & 0 & 0 \\ .5 & .5 & 0 & 0 \\ 0 & 0 & .1 & .9 \\ 0 & 0 & .1 & .9 \end{pmatrix}.$$

Then, if the two chains start from state 3 or 4, then it is possible to distinguish between them; however, if they start from state 1 or 2, then the resulting distributions are the same.

Now we are ready to show that \bar{D}_m decays exponentially.

Theorem 2 *For every block stationary block Markov source X_0^∞ there is a constant $c > 0$ depending on the transition matrix of the source such that*

$$\limsup_{m \rightarrow \infty} \frac{1}{m} \log \bar{D}_m \leq -c.$$

Proof. First notice that $U_0^\infty = \{U_{2kN}^{2(k+1)N-1}\}_{k=0}^\infty$ is a block- $2N$ Markov source for each value of τ , as $U_{2kN}^{2(k+1)N-1}$ always contains a full character of the block- N Markov source $\{X_{jN}^{(j+1)N-1}\}_{j=0}^\infty$.

This fact will enable us to use tools for classification of Markov chains (namely, Lemma 2) to examine estimates of τ based on the sequence U_0^{m-1} , which then can be used to estimate $\bar{D}_m = I(\tau; U_m | U_0^{m-1})$.

For $\tau = t, t \in \{0, \dots, N-1\}$ and $U_0^{2N-1} = w \in \mathcal{A}^{2N}$, let $\mathcal{I}_{t,w} \subset \mathcal{A}^{2N}$ denote the (irreducible) set of states reachable from w by the Markov chain. Moreover, let $Q_{t,w} = \{q_t(v|u)\}, u, v \in \mathcal{I}_{t,w}$ denote the transition probability matrix corresponding to the states in $\mathcal{I}_{t,w}$. That is,

$$q_t(u|v) = P_{U_{2N}^{4N-1} | U_0^{2N}, \tau}(u|v, t) = P_{X_{2N+t-m}^{4N+t-m-1} | X_{t-m}^{2N+t-m-1}}(v|u).$$

To simplify further notation, we extend the above definition for any integer t . Note that $Q_{t,w}$ is a sub-matrix of the transition probability matrix $\{q_t(v|u)\}, u, v \in \mathcal{A}^{2N}$ describing the behavior of the Markov chain for all states. Moreover, $Q_{t,w}$ is irreducible for all w , but the index set $\mathcal{I}_{t,w}$ is not necessarily the same for different values of t . If $\mathcal{I}_{t,w} \neq \mathcal{I}_{t',w}$, then the corresponding matrices $Q_{t,w}$ and $Q_{t',w}$ are different. However, $Q_{kN+t,w} = Q_{t,w}$ for any integer $k, t \in \{0, \dots, N-1\}$ and $w \in \mathcal{A}^{2N}$, since $X_{-\infty}^{\infty}$ is block- N stationary.

In the proof we will try to estimate which $Q_{t,w}$ is the generator matrix of an observed sequence U_0^m . Obviously, if the $Q_{t,w}$ are not all different, this is not possible (we cannot distinguish between two Markov-chains with the same transition matrices). Therefore, for any t let $g_w(t)$ denote the smallest number in $\{0, \dots, N-1\}$ such that $Q_{t,w} = Q_{g_w(t),w}$, and let N_w^* be the number of different transition matrices $Q_{t,w}$. (It is easy to show that $N_w^* = \max_{0 \leq t < N} g_w(t) + 1$, and $Q_{0,w}, \dots, Q_{N_w^*-1,w}$ are different.)

Moreover, given $U_0^{2N-1} = w$ and $g_w(\tau)$, τ is independent of U_0^m for every m . Therefore,

$$\begin{aligned} H(\tau | U_{2N}^m, U_0^{2N-1} = w) &= H(\tau, g_w(\tau) | U_{2N}^m, U_0^{2N-1} = w) \\ &= H(g_w(\tau) | U_{2N}^m, U_0^{2N-1} = w) + H(\tau | g_w(\tau), U_{2N}^m, U_0^{2N-1} = w) \\ &= H(g_w(\tau) | U_{2N}^m, U_0^{2N-1} = w) + H(\tau | g_w(\tau)). \end{aligned}$$

By Theorem 1, this implies for all $m \geq 2N$

$$\begin{aligned} \bar{D}_m &= I(\tau; U_m | U_0^{m-1}) = H(\tau | U_0^{m-1}) - H(\tau | U_0^m) \\ &= \sum_w P_{U_0^{2N-1}}(w) (H(g_w(\tau) | U_{2N}^{m-1}, U_0^{2N-1} = w) - H(g_w(\tau) | U_{2N}^m, U_0^{2N-1} = w)) \\ &\leq \sum_w P_{U_0^{2N-1}}(w) H(g_w(\tau) | U_{2N}^{m-1}, U_0^{2N-1} = w) \\ &\leq \max_{w: P_{U_0^{2N-1}}(w) > 0} H(g_w(\tau) | U_{2N}^{m-1}, U_0^{2N-1} = w). \end{aligned}$$

Therefore,

$$\limsup_{m \rightarrow \infty} \frac{1}{m} \log \bar{D}_m \leq \max_{w: P_{U_0^{2N-1}}(w) > 0} \limsup_{m \rightarrow \infty} \frac{1}{m} \log H(g_w(\tau) | U_{2N}^{m-1}, U_0^{2N-1} = w). \quad (1)$$

Next we bound the conditional entropies $H(g_w(\tau)|U_{2N}^{m-1}, U_0^{2N-1} = w)$. If $N_w^* = 1$, then $g_w(\tau) = 0$ with probability 1, and so $H(g_w(\tau)|U_{2N}^{m-1}, U_0^{2N-1} = w) = 0$ and

$$\lim_{m \rightarrow \infty} \frac{1}{m} \log H(g_w(\tau)|U_{2N}^{m-1}, U_0^{2N-1} = w) \leq -c_w \quad (2)$$

with $c_w = \infty$. Otherwise, if $N_w^* > 1$, let $\tau_{m,w} = \tau_{m,w}(U_0^{m-1})$ be an optimal estimate of $g_w(\tau)$ based on U_0^{m-1} , given $U_0^{2N-1} = w$ in the sense that $\Pr(g_w(\tau) = \tau_{m,w}) \geq \Pr(g_w(\tau) = f(U_0^{m-1})|U_0^{2N-1} = w)$ for any function $f: \mathcal{A}^m \rightarrow \{0, \dots, N^* - 1\}$, and let

$$p_{m,w} = \Pr(g_w(\tau) \neq \tau_{m,w}|U_0^{2N-1} = w)$$

(note that such an estimate always exist). Then, since $\tau_{m,w}$ is a function of U_0^m ,

$$H(g_w(\tau)|U_{2N}^{m-1}, U_0^{2N-1} = w) \leq H(g_w(\tau)|\tau_{m,w}, U_0^{2N-1} = w) \quad (3)$$

(for properties of the entropy function see, e.g., [5]). Moreover, by Fano's inequality

$$H(\tau^*|\tau_m, U_0^{2N-1} = w) \leq \log(N_w^* - 1)p_{m,w} + h_b(p_{m,w})$$

where $h_b(p) = -p \log p - (1-p) \log(1-p)$ for $0 \leq p \leq 1$. From here, obviously

$$\begin{aligned} & \limsup_{m \rightarrow \infty} \frac{1}{m} \log H(g_w(\tau)|\tau_{m,w}, U_0^{2N-1} = w) \\ & \leq \limsup_{m \rightarrow \infty} \frac{1}{m} \log (2 \max\{\log(N_w^* - 1)p_{m,w}, h_b(p_{m,w})\}) \\ & \leq \max \left\{ \limsup_{m \rightarrow \infty} \frac{1}{m} \log p_{m,w}, \limsup_{m \rightarrow \infty} \frac{1}{m} \log h_b(p_{m,w}) \right\}. \end{aligned} \quad (4)$$

Next we use Lemma 2 to bound (4). In order to be able to apply the lemma, we need to determine the state space of the observed process, and then we only need to find the generating Markov chain (given by $Q_{t,w}$) among those that live on that state space. Since the Markov chains defined by the matrices $Q_{t,w}$ are irreducible, the probability that a given state is not reached in k steps converges to 0 exponentially fast in k . Therefore, for the set of values $\hat{\mathcal{I}}_k = \{U_0^{2N-1}, U_{2N}^{4N-1}, \dots, U_{2(k-1)N}^{2kN-1}\}$, we have

$$\lim_{k \rightarrow \infty} \Pr(\hat{\mathcal{I}}_k \neq \mathcal{I}_{t,w} | \tau = t, U_0^{2N-1} = w) \leq -c'_{w,t}$$

for some $c'_{w,t} > 0$. This implies that

$$\lim_{m \rightarrow \infty} \Pr(\hat{\mathcal{I}}_{\lfloor \frac{m}{2N} \rfloor} \neq \mathcal{I}_{g_w(\tau)} | U_0^{2N-1} = w) \leq -c_{w,1} \quad (5)$$

where $c_{w,1} = \min_t c'_{w,t}/2N > 0$.

For any $\mathcal{I} \subset \mathcal{A}^{2N}$, let

$$\mathcal{Q}_w(\mathcal{I}) = \{g_w(i) : 0 \leq i < N, Q_{i,w} \text{ is defined on } \mathcal{I}\}$$

denote the set of indexes of the Markov chains with state space \mathcal{I} (for Markov chains with the same transition matrix, we consider the one with the smallest index). Now $g_w(\tau)$ can be estimated by first estimating $\mathcal{I}_{g_w(\tau)}$ by $\hat{\mathcal{I}}_k$, $k = \lfloor (m-1)/2N \rfloor$, based on U_0^m , and then estimating $g_w(\tau)$ by the optimal classifier for the problem of deciding which $Q_{i,w}$, $i \in \mathcal{Q}_w(\hat{\mathcal{I}}_k)$ generated the sequence $U_0^{2N-1} = w, U_{2N}^{4N-1}, \dots, U_{2(k-1)N}^{2kN-1}$. Let $p'_{m,w}$ denote the conditional error probability of the optimal classifier $\hat{\mathcal{I}}_k$ for $\mathcal{I}_{g_w(\tau)}$ given $U_0^{2N-1} = w$. Then

$$p_{m,w} \leq \Pr(\hat{\mathcal{I}}_k \neq \mathcal{I}_{g_w(\tau)} | U_0^{2N-1} = w) + p'_{m,w}. \quad (6)$$

Let

$$R_i = \{P \in \Lambda : D(P||Q_{i,w}) < D(P||Q_{j,w}) \text{ for all } i \neq j, i, j \in \mathcal{Q}_w(\mathcal{I}_{g_w(\tau)})\}$$

and define

$$c_{w,2} = \min_{i \in \mathcal{Q}_w(\mathcal{I}_{g_w(\tau)})} \min_{P \in \Lambda_0 \cap \bar{R}_i} D(P||Q_{i,w}).$$

Then, as the $\{Q_i\}, i \in \mathcal{Q}_w(\mathcal{I}_{g_w(\tau)})$ are different and irreducible, from Lemma 2 we have $c_{w,2} > 0$ and

$$\limsup_{m \rightarrow \infty} \frac{1}{m} \log p'_{m,w} \leq -c_{w,2}.$$

Combining this inequality with (5) and (6) we obtain that for the positive number $c_w = \min\{c_{w,1}, c_{w,2}\}$ we have

$$\limsup_{m \rightarrow \infty} \frac{1}{m} \log p_{m,w} \leq -c_w. \quad (7)$$

In particular, $\lim_{m \rightarrow \infty} p_{m,w} = 0$. Therefore, as L'Hospital's rule implies

$$\lim_{p \rightarrow 0} p \log(1/p) / h_b(p) = 1$$

we have

$$\begin{aligned} \limsup_{m \rightarrow \infty} \frac{1}{m} \log h_b(p_{m,w}) &= \limsup_{m \rightarrow \infty} \frac{1}{m} \log \left(p_{m,w} \log \frac{1}{p_{m,w}} \right) \\ &= \limsup_{m \rightarrow \infty} \frac{1}{m} \log p_{m,w} \end{aligned}$$

where the second equality holds because (7) implies $\lim_{m \rightarrow \infty} \frac{1}{m} \log \log p_{m,w} = 0$. Thus, the two terms in the maximum in (4) are equal and converge to zero exponentially fast by (7). Combining this fact with inequalities (1) and (2) proves the theorem. \square

5 Universal symbol-based coding of block Markov sources

Now we are ready to establish an upper bound for the real coding redundancy for a large class of universal symbol-based codes. Let $\ell_n^{(m)} : \mathcal{A}^n \rightarrow \{0, 1\}^*$ denote the code lengths of a universal code $\{f_n\}$ for m th order Markov sources satisfying

$$\frac{1}{n} \sup_{P_{Y_0^{n-1}}} \sup_{z_0^{n-1} \in \mathcal{A}^n} \left[\ell_n^{(m)}(z_0^{n-1}) + \log P_{Y_0^{n-1}}(z_0^{n-1}) \right] \leq c_n^{(m)} \quad (8)$$

for some $c_n^{(m)} \rightarrow 0$ as $n \rightarrow \infty$, where the first supremum is taken over all n -fold marginal distributions of m th order Markov sources over \mathcal{A} . In other words, we require that the ‘‘pointwise redundancy’’ converges to zero uniformly for each source sequence and for each m th order Markov source. For example, there exist universal arithmetic codes for m th order Markov sources with $c_n^{(m)} = O(|\mathcal{A}|^{m+1} \log n/n)$ (see, e.g., [6]).

For fixed m and n the per symbol coding redundancy is defined as

$$R_{n,m} = \frac{1}{n} (E \ell_n^{(m)}(X_0^{n-1}) - H(X_0^{n-1}))$$

Theorem 3 *If the code length function $\ell_n^{(m)}$ satisfies (8) then for $n \geq m \geq 2N$ the coding redundancy $R_{n,m}$ for the block stationary block Markov source X_0^∞ can be bounded as*

$$R_{n,m} \leq \frac{1}{n} \log N + 2^{-mc_r + o(m)} + c_n^{(m)} \quad (9)$$

where c_r is defined in Theorem 2.

Remarks.

(i) For any fixed m and very large coding block length n , the redundancy is exponentially small in m , that is,

$$\limsup_{n \rightarrow \infty} \frac{1}{n} E[\ell_n^{(m)}(X_0^{n-1})] - \bar{H}(X_0^\infty) \leq 2^{-mc_r + o(m)}.$$

(ii) It is easy to see that to minimize the bound (9), m should be chosen $O(\log n)$. As mentioned before, there are arithmetic codes with $c_n^{(m)} = O(|\mathcal{A}|^{m+1} \log n/n)$ [6]. For these codes, the optimal choice is $m = \log n / (c_r + \log |\mathcal{A}|)$, yielding a redundancy bound of order $n^{-\frac{c_r}{c_r + \log |\mathcal{A}|}}$. Obviously, c_r is not known in advance. Moreover, this rate is slower than applying the universal code to the first order block Markov source, which results in $O(N|\mathcal{A}|^{2N} \log(n/N)/n)$ redundancy. The reason for this is that while the number of parameters of the original source is finite (namely, $O(|\mathcal{A}|^{2N})$), the number of parameters of the approximating m th order Markov chain (which is $O(|\mathcal{A}|^m)$) grows without bound as m increases. On the other hand, if the dependence of $c_n^{(m)}$ on m is less than exponential, then the dominant term in (9) is usually the last one.

(iii) The result may be interesting for the practical case of universal compression, when the block size of the input is not known. Choosing an incorrect block length may result in deteriorated performance, as illustrated by the following experiment. We used two representations of the English-language text “book1” from the Calgary Corpus [3], one using $N_s = 7$ bits per character, the other using $N_s = 8$ bits per character. The resulting files were compressed with the `bzip2` algorithm operating on (possibly different) fixed-length blocks of N_e symbols (N_e is chosen to be 1,7, and 8).¹ Obviously, the plots when $N_s = N_e$ are the same. The per-block entropy rate of the source does not depend on N_s , equaling approximately 2 bits per block. The graph in Fig. 1 shows the average number of bits in the encodings per one source block, as the length of the source sequence (measured in source blocks) increases. It can be seen that the performance for the the binary alphabet size ($N_e = 1$) is robust and superior to the case where the source and encoder alphabet sizes are mismatched (i.e, either $N_s = 7$ and $N_e = 8$, or $N_s = 8$ and $N_e = 7$).

The experiment is repeated with a truly first-order Markov source which was generated from “book1” using the text’s empirical first-order Markov transition probabilities. Fig. 2 shows the results which are consistent with that of the first experiment. Here we know that choosing the smallest encoding block length $N_e = 1$ results in guaranteed performance by Theorem 3, with the computational advantage of operating on a small alphabet. Thus, coding on the elementary symbol level is a practically good suboptimal scheme for encoding block Markov sources with unknown block size.

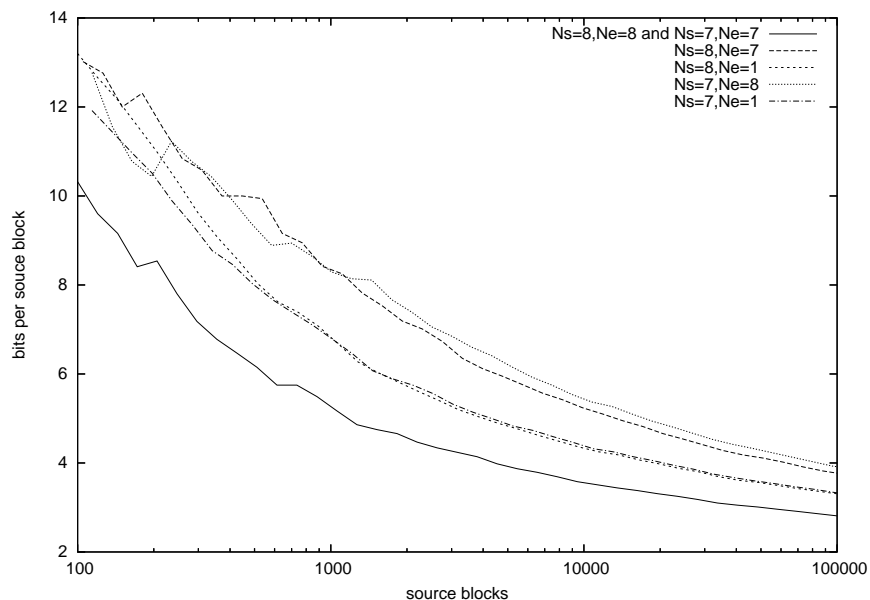


Figure 1: Compression performance of `bzip2` on “book1” for different source and encoder block sizes.

¹Of course, strictly speaking `bzip2` is not a universal compression method, but it serves well for illustrative purposes.

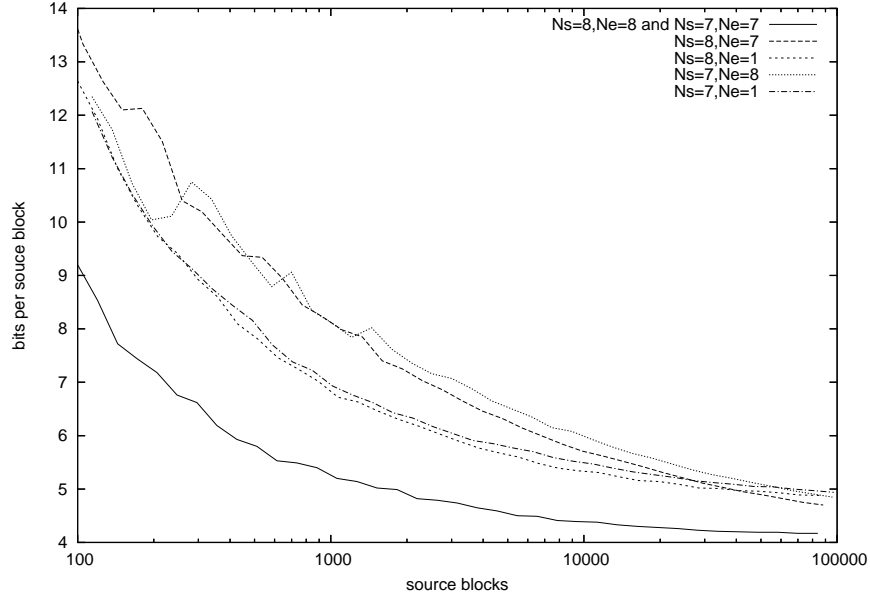


Figure 2: Compression performance of bzip2 on the first-order Markov source generated from ‘book1’ for different source and encoder block sizes.

Proof of Theorem 3: From (8) it follows that for any m th order Markov source Y_0^∞ and x_0^{n-1}

$$\log \frac{P_{Y_0^{n-1}}(x_0^{n-1})}{P_{\ell_n^{(m)}}(x_0^{n-1})} \leq nc_n^{(m)} \quad (10)$$

where $P_{\ell_n^{(m)}}$ denotes the coding distribution for n -long sequences.

Let \hat{Y}_0^∞ denote the stationary m th order Markov approximation of X_0^∞ , defined in Theorem 1, achieving the minimum in the definition of \bar{D}_m (recall that $P_{\hat{Y}_0^{m-1}} = P_{U_0^{m-1}}$). Then

$$\begin{aligned} & D(P_{X_0^{n-1}} \| P_{\ell_n^{(m)}}) \\ &= D(P_{X_0^{n-1}} \| P_{\hat{Y}_0^{n-1}}) + \sum_{z_0^{n-1} \in \mathcal{A}^n} P_{X_0^{n-1}}(z_0^{n-1}) \log \frac{P_{\hat{Y}_0^{n-1}}(z_0^{n-1})}{P_{\ell_n^{(m)}}(z_0^{n-1})} \\ &\leq D(P_{X_0^{n-1}} \| P_{\hat{Y}_0^{n-1}}) + nc_n^{(m)} \end{aligned}$$

where the inequality holds by (10). Now, the first term can be easily bounded following the proof

of Theorem 1 as

$$\begin{aligned}
D(P_{X_0^{n-1}} \| P_{\hat{Y}_0^{n-1}}) &= D(P_{X_0^{m-1}} \| P_{\hat{Y}_0^{m-1}}) + \sum_{i=m}^n D(P_{X_i|X_0^{i-1}} \| P_{Y_i|Y_0^{i-1}}) \\
&\leq D(P_{X_0^{m-1}} \| P_{\hat{Y}_0^{m-1}}) + \sum_{i=m}^{m-1+N \lceil \frac{n-m+1}{N} \rceil} D(P_{X_i|X_0^{i-1}} \| P_{Y_i|Y_0^{i-1}}) \\
&= D(P_{X_0^{m-1}} \| P_{\hat{Y}_0^{m-1}}) + \left\lceil \frac{n-m+1}{N} \right\rceil \sum_{t=0}^{N-1} S_t \\
&\leq D(P_{X_0^{m-1}} \| P_{\hat{Y}_0^{m-1}}) + n\bar{D}_m
\end{aligned}$$

where S_t is defined as in the proof of Theorem 1 with $Y_0^\infty = \hat{Y}_0^\infty$. Furthermore,

$$D(P_{X_0^{m-1}} \| P_{\hat{Y}_0^{m-1}}) = D(P_{X_0^{m-1}} \| P_{U_0^{m-1}}) \leq \log N$$

since for any $x_0^{m-1} \in \mathcal{A}^m$, $P_{U_0^{m-1}}(x_0^{m-1}) \geq P_{X_0^{m-1}}(x_0^{m-1})/N$ by definition. Thus, by Theorem 2

$$R_{n,m} \leq \frac{1}{n} \log N + 2^{-mc_r + o(m)} + c_n^{(m)}.$$

□

6 Conclusion

We have demonstrated that block Markov sources can be encoded with exponentially fast vanishing redundancy using codes that are optimized for higher-order symbol-level Markov models. This partially explains the findings of our experiments that a bit-level implementation of a universal compression algorithm performs reasonably well on byte-aligned data when compared with byte-level implementations, inviting further studies of bit-level implementations of compression algorithms, as on the bit level, one can take advantage of the computational benefits of operating on the smallest possible alphabet.

References

- [1] D. Nagy and T. Linder, “Experimental study of a binary block sorting compression scheme,” in *Proceedings of Data Compression Conference, DCC’03*, (Snowbird, UT, USA), p. 439, IEEE Computer Society Press, Apr. 2003.
- [2] M. Burrows and D. J. Wheeler, “A Block-sorting lossless data compression algorithm,” *SRC Research Report 124*, DSRC, Palo Alto, CA, May 1994.

- [3] I. Witten and T. Bell, *The Calgary Text Compression Corpus*, available via anonymous ftp at <ftp.cpcs.ucalgary.ca/pub/projects/text.compression.corpus>
- [4] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.
- [5] T. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [6] I. Csiszár and P. C. Shields, "Information Theory and Statistics: A Tutorial", *Foundations and Trends in Communications and Information Theory*, vol. 1, issue 4, Dec. 2004.
- [7] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.
- [8] S. Natarajan, "Large deviations, hypothesis testing, and source coding for finite Markov chains," *IEEE Trans. Information Theory*, vol. IT-31, pp. 360–365, May 1985.
- [9] V. Anantharam, "A large deviations approach to error exponents in source coding and hypothesis testing," *IEEE Trans. Information Theory*, vol. IT-36, pp. 938–943, July 1990.
- [10] N. Merhav and J. Ziv, "A Bayesian approach for classification of Markov sources," *IEEE Trans. Information Theory*, vol. IT-37, pp. 1067–1071, July 1991.
- [11] I. Csiszár, T. M. Cover, and B.-S. Choi, "Conditional limit theorems under Markov conditioning," *IEEE Trans. Information Theory*, vol. IT-33, pp. 788–801, Nov. 1987.