Symbol-Based Modeling and Coding of Block Markov Sources

Dániel A. Nagy, András György, *Member, IEEE*, and Tamás Linder, *Senior Member, IEEE*

***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 correspondence, 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 (our model for byte-based data) with higher order Markov sources (which model bit-based Markov encoders), 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*—Binary codes, block Markov sources, byte-aligned sources, higher order Markov modeling, lossless coding.**

I. 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 (8-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 three 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 suboptimal bit-level Burrows–Wheeler transform (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 `bzip2`, a similar compression algorithm implemented on bytes.

Motivated by these practical observations, in this correspondence we evaluate this penalty in an information-theoretic setting. Specifically, under Markovian assumptions we investigate the excess of the 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 this correspondence, we use a block Markov source to model data whose "natural" symbols are blocks of a given length (e.g., bytes) from a source alphabet of elementary symbols (e.g., bits). An encoder that operates on elementary symbols (or more precisely, the associated model distribution) is identified with the distribution of a higher order Markov source.

In Section III, 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 IV, 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 V, 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.

II. Preliminaries

For any sequence of random variables

$$\{X_n\}_{n=0}^{\infty} = X_0, X_1, \ldots, X_n, \ldots$$

and for any $i \le j$, the segment $(X_i, X_{i+1}, \ldots, X_j)$ 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

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

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

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

provided the limit exists.

Manuscript received September 2, 2005; revised August 24, 2006. This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Hungarian Scientific Research Fund (OTKA F60787), and the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. The material in this correspondence was presented in part at the 2005 Canadian Workshop on Information Theory, Montreal, QC, Canada, June 2005.

D. A. Nagy was with the Department of Mathematics and Statistics, Queen's University, Kingston, ON K7L 3N6, Canada. He is now with the Department of Computer Science, Eötvös Lóránd University, 1117 Budapest, Hungary (e-mail: nagydani@cs.elte.hu).

A. György is with the Machine Learning Research Group, Computer and Automation Research Institute of the Hungarian Academy of Sciences, 1111 Budapest, Hungary (e-mail: gya@szit.bme.hu).

T. Linder is with the Department of Mathematics and Statistics, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: linder@mast.queensu.ca).

Communicated by W. Szpankowski, Associate Editor for Source Coding.

Digital Object Identifier 10.1109/TIT.2006.885539

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 $x_0^{(i+1)N-1} \in \mathcal{A}^{(i+1)N}$

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

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

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

A binary block code of length n for the source alphabet $\mathcal{A}$ is given by a function $f_n : \mathcal{A}^n \to \{0,1\}^*$, which maps any source vector $x \in \mathcal{A}^n$ to the binary string $f_n(x)$. The length function $\ell_n : \mathcal{A}^n \to \mathbb{N}$ 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 $x_1, \ldots, x_j, y_1, \ldots, y_k \in \mathcal{A}^n$

$$f_n(x_1)f_n(x_2)\ldots f_n(x_j) = f_n(y_1)f_n(y_2)\ldots f_n(y_k)$$

if and only if $j = k$ and $x_i = y_i, i = 1, \ldots, j$, where for two binary strings s_1 and s_2, $s_1 s_2$ denotes their concatenation. It is well known [5] that if f_n is uniquely decodable than its length function ℓ satisfies the Kraft inequality

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

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 $\ell_n'(x) \le \ell_n(x)$ 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 remainder of the correspondence, 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

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

the difference of the expected code length $E\ell_n(X_1^n)$ and the entropy

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

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

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

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

$$R_n' = E\ell_n'\left(X_0^{n-1}\right) - H\left(X_0^{n-1}\right) \le D\left(X_0^{n-1} \| \hat{Y}_0^{n-1}\right) + 1$$

where $\hat{Y}_0^{n-1}$ 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

$$\lim_{n\to\infty} \frac{1}{n} R_n = \lim_{n\to\infty} \frac{1}{n} D\left(X_0^{n-1} \| Y_0^{n-1}\right) = \bar{D}\left(X_0^\infty \| Y_0^\infty\right)$$

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

In the sequel, depending on the context, a *code* will either mean a block-n code f_n, or a sequence of such codes $\{f_n\}_{n=1}^\infty$.

III. Approximation of Block Markov Sources

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

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

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

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

where τ is a random variable that is uniformly distributed on $\{0, 1, \ldots, N-1\}$ 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 mth-order Markovian approximation of $\{U_n\}$, which has the same mth-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 mth-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 $\hat{Y}_0^\infty$ of Y_0^∞ such that $P_{\hat{Y}_0^m} = P_{U_0^m}$.

Expressing conditional mutual information in terms of conditional entropies as

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

we obtain

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

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

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

Remark 1: The fact that $\bar{D}_m$ converges to zero as $m \to \infty$ is not very surprising in view of the fact that the divergence rate between a stationary process and its best mth-oder Markov approximation asymptotically vanishes as $m \to \infty$ (see, e.g., [7]). Note, however, that X_0^∞ is nonstationary, 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: First note that $\min \bar{D}(X_0^\infty \| Y_0^\infty)$ is finite (for example, if the Y_n are independent and identically distributed according to the one-dimensional marginal distribution of X_0^∞). Therefore, to find the minimum, it is enough to consider $\{Y_n\}$ sequences such that $D(X_0^n \| Y_0^n) = D(P_{X_0^n} \| P_{Y_0^n})$ is finite for all n.

For all $n > m$, we have from the chain rule for the relative entropy [5]

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

where

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

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

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

and

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

Therefore

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

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

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n+1} D\left(P_{X_0^n} \| P_{Y_0^n}\right) \\
&\qquad = \lim_{n\to\infty} \frac{1}{n+1} \sum_{i=m}^{n} D\left(P_{X_i|X_0^{i-1}} \| P_{Y_i|Y_0^{i-1}}\right) \\
&\qquad = \frac{1}{N} \sum_{t=0}^{N-1} S_t.
\end{aligned}$$

Let τ denote a uniform random variable over $\{0, 1, \ldots N-1\}$ that is independent of the pair $(\{X_n\}, \{Y_n\})$, 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

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

Observe that only the second term of the preceding 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

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

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 mth-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. □

From a coding point of view, Theorem 1 states that if a sequence of block codes is optimal for Y_0^∞ (in the sense that the length function of the nth code is matched to the nth-order marginal distribution of Y_0^∞), then it can asymptotically compress X_0^∞ with rate not exceeding the source entropy rate $\bar{H}(X_0^\infty) = \lim_{n-\infty} H(X_0^{n-1})$ 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 only *almost* optimal for Y_0^∞ in the expected codeword length sense need not be good at all for X_0^∞. Still, it is reasonable to expect that codes that are universal for the class of mth-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 V.

IV. Rate of Convergence

In this section, we examine the rate at which the minimum relative entropy rate $\bar{D}_m$ converges to zero 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 higher 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]–[10]. However, the conditions imposed in these works are not immediately applicable to our setup. We will use an approach based on Bayesian hypothesis testing following the lines of the derivation of the Chernoff bound (see, e.g., [5]) to obtain an upper bound on the classification error under more general conditions. Another approach, which gives somewhat less explicit results, is to combine results from Csiszár *et al.* [11] on large deviations for Markov chains with the method used by Natarayan [8] (this approach was used in earlier versions of this work [12] and [13]).

Assume that the sample $X_1, X_2, \ldots, X_n$ is generated by one of K stationary Markov sources over a finite alphabet $\mathcal{A}$ with transition matrices $W_1, \ldots, W_K$. 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 entry in the (u,v) position of W^n is positive.)

Lemma 1: Let $\{X_{i,n}\}_{n=0}^\infty$, $i = 1, \ldots, K, K \geq 2$ 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 $\{1, \ldots, K\}$ such that $\Pr(t = i) > 0$ for all $i = 1, \ldots, K$, and t is independent of the $\{X_{i,n}\}$'s. Finally, assume that we observe the tth Markov source, that is, let $X_n = X_{t,n}$ for $n = 0, 1, \ldots$ Define

$$R_i^{(n)} = \left\{ x_0^n \in \mathcal{A}^{n+1} : \prod_{k=1}^n W_i(x_k|x_{k-1}) > \prod_{k=1}^n W_j(x_k|x_{k-1}) \text{ for all } j \neq i \right\}$$

and let $\hat{t}_n = i$ if $X_0^n \in R_i^{(n)}$ for some $i \in \{1, \ldots, K\}$ and let $\hat{t}_n$ be arbitrary otherwise.

Then for any $u \in \mathcal{A}$

$$\limsup_{n-\infty} \frac{1}{n} \log \Pr(t \neq \hat{t}_n | X_0 = u) \leq \max_{1\leq i\neq j\leq K} \min_{0\leq\lambda\leq 1} \log \rho(W_{i,j,\lambda}) < 0 \quad (1)$$

where the entry in the (u,v) position of the matrix $W_{i,j,\lambda}$ is given by

$$W_{i,j,\lambda}(v|u) = W_i(v|u)^\lambda W_j(v|u)^{1-\lambda}$$

for any $u, v \in \mathcal{A}$, $\rho(W) = \max\{|\alpha| : \alpha \text{ is an eigenvalue of } W\}$ is the spectral radius of a square matrix W, and $\log 0 = -\infty$ by definition.

Remark: Note that for any $1 \leq i \neq j \leq K$, the minimum over λ in (1) can be computed numerically.

Proof of Lemma 1: First we show, by adapting the proof of the Chernoff bound in [5], that

$$\limsup_{n-\infty} \frac{1}{n} \log \Pr(t \neq \hat{t}_n | X_0 = u) \leq \max_{1\leq i\neq j\leq K} \min_{0\leq\lambda\leq 1} \log \rho(W_{i,j,\lambda}).$$

Then, to conclude the proof, we prove that $\min_{0\leq\lambda\leq 1} \rho(W_{i,j,\lambda}) < 1$ for all $i \neq j$. (Note that the minimum exists as $\rho(W_{i,j,\lambda})$ is a continuous function of λ.)

Extend the regions $R_i^{(n)}$ to their boundaries by defining, for all $1 \leq i \leq K$

$$\bar{R}_i^{(n)} = \left\{ x_0^n \in \mathcal{A}^{n+1} : \prod_{k=1}^n W_i(x_k|x_{k-1}) \geq \prod_{k=1}^n W_j(x_k|x_{k-1}) \text{ for all } j \neq i \right\}.$$

Then, for any $0 \le \lambda \le 1$,

$$\begin{aligned}
&\Pr(t \neq \hat{t}_n | X_0 = u) \\
&\quad = \sum_{i=1}^{K} \Pr(t=i) \Pr(\hat{t}_n \neq i | X_0 = u, t = i) \\
&\quad \le \sum_{i=1}^{K} \Pr(t=i) \sum_{x_0^n \notin R_i^{(n)}, x_0 = u} \prod_{k=1}^{n} W_i(x_k|x_{k-1}) \\
&\quad \le \sum_{i=1}^{K} \Pr(t=i) \sum_{j \neq i} \sum_{x_0^n \in \bar{R}_j^{(n)}, x_0 = u} \prod_{k=1}^{n} W_i(x_k|x_{k-1}) \\
&\quad = \sum_{i=1}^{K} \Pr(t=i) \sum_{j \neq i} \sum_{x_0^n \in \bar{R}_j^{(n)}, x_0 = u} \\
&\qquad \min\left\{ \prod_{k=1}^{n} W_i(x_k|x_{k-1}), \prod_{k=1}^{n} W_j(x_k|x_{k-1}) \right\} \\
&\quad \le \sum_{i=1}^{K} \Pr(t=i) \sum_{j \neq i} \sum_{x_0^n \in \bar{R}_j^{(n)}, x_0 = u} \\
&\qquad \left(\prod_{k=1}^{n} W_i(x_k|x_{k-1}) \right)^{\lambda} \left(\prod_{k=1}^{n} W_j(x_k|x_{k-1}) \right)^{1-\lambda} \qquad (2) \\
&\quad = \sum_{i=1}^{K} \Pr(t=i) \sum_{j \neq i} \sum_{x_0^n \in \bar{R}_j^{(n)}, x_0 = u} \prod_{k=1}^{n} W_i(x_k|x_{k-1})^{\lambda} \\
&\qquad \times W_j(x_k|x_{k-1})^{1-\lambda} \\
&\quad \le \sum_{i=1}^{K} \Pr(t=i) \sum_{j \neq i} \sum_{x_1^n \in \mathcal{A}^n, x_0 = u} \prod_{k=1}^{n} W_i(x_k|x_{k-1})^{\lambda} \\
&\qquad \times W_j(x_k|x_{k-1})^{1-\lambda} \\
&\quad = \sum_{i=1}^{K} \Pr(t=i) \sum_{j \neq i} e_u^T W_{i,j,\lambda}^n e \qquad (3)
\end{aligned}$$

where $e_u^T = (0, \ldots, 0, 1, 0, \ldots, 0)$ is a unit row vector indexed with the elements of $\mathcal{A}$ with a 1 at position u and zero elsewhere, and e is a column vector whose entries are all 1's. The inequality in (2) holds since for any $p, q \ge 0$ and $0 \le \lambda \le 1$, $\min\{p, q\} \le p^{\lambda} q^{1-\lambda}$.

Let $\|W\|_1$ denote the l_1 norm (sum of the absolute values of all entries) of a square matrix W. It is known [14, Corollary 5.6.14] that $\lim_{n\to\infty} (\|W^n\|_1)^{1/n} = \rho(W)$. Since $W_{i,j,\lambda}^n$ is a nonnegative matrix, we have that

$$e_u^T W_{i,j,\lambda}^n e \le e^T W_{i,j,\lambda}^n e = \|W_{i,j,\lambda}^n\|_1$$

so that

$$\limsup_{n\to\infty} \frac{1}{n} \log e_u^T W_{i,j,\lambda}^n e \le \log \rho(W_{i,j,\lambda}). \qquad (4)$$

This and (3) imply the first inequality in (1).

To show that $\min_{0 \le \lambda \le 1} \rho(W_{i,j,\lambda}) < 1$ for all $i \neq j$, we use the fact that the spectral radius of a nonnegative matrix is bounded by the maximum row sum of the matrix [14, Theorem 8.1.22]. Thus, for any $0 \le \lambda \le 1$

$$\begin{aligned}
\rho(W_{i,j,\lambda}) &\le \max_{u \in \mathcal{A}} \sum_{v \in \mathcal{A}} W_{i,j,\lambda}(v|u) \qquad (5) \\
&= \max_{u \in \mathcal{A}} \sum_{v \in \mathcal{A}} W_i(v|u)^{\lambda} W_j(v|u)^{1-\lambda} \\
&\le \max_{u \in \mathcal{A}} \sum_{v \in \mathcal{A}} \lambda W_i(v|u) + (1-\lambda) W_j(v|u) \qquad (6) \\
&= 1 \qquad (7)
\end{aligned}$$

where (7) holds since W_i and W_j are transition matrices with unit row sums, and (6) follows from the well-known inequality $p^{\lambda} q^{1-\lambda} \le \lambda p + (1-\lambda) q$ for $p, q \ge 0, 0 \le \lambda \le 1$, where equality holds if and only if (iff) $p = q$ or $\lambda \in \{0, 1\}$. We show that equality is not possible in (5) and (6) simultaneously if $\lambda \in (0, 1)$.

Let $\lambda \in (0, 1)$ and assume (5) and (6) are both equalities so that $\rho(W_{i,j,\lambda}) = 1$. If $W_{i,j,\lambda}$ is irreducible, then equality holds in the upper bound (5) iff all rows of $W_{i,j,\lambda}$ sum to $\rho(W_{i,j,\lambda})$ [15, p. 287, Exercise 4]. In general, $W_{i,j,\lambda}$ may not be irreducible, but using the irreducible normal form of $W_{i,j,\lambda}$, we have that equality holds in (5) iff there is a set $\hat{\mathcal{A}} \subset \mathcal{A}$ such that the submatrix $\hat{W}$ composed of the entries of $W_{i,j,\lambda}$ indexed by $\hat{\mathcal{A}}^2$ is irreducible, and its row sums are equal to $\rho(W_{i,j,\lambda})$. Then, repeating the derivation (5)–(7) with $\hat{W}$ in place of $W_{i,j,\lambda}$, if equality holds throughout the derivation, then $W_i(v|u) = W_j(v|u) = \hat{W}(v|u)$ for all $u, v \in \hat{\mathcal{A}}$, and so $\hat{W}$ is a transition probability matrix. Therefore, since W_i and W_j are also transition matrices, and all three matrices are irreducible, we obtain $\hat{\mathcal{A}} = \mathcal{A}$, and hence $W_i = W_j = \hat{W}$, which contradicts the assumption of the lemma. Thus, we must have $\rho(W_{i,j,\lambda}) < 1$, which proves the second inequality in (1). □.

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:

$$W_1 = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 0.5 \end{pmatrix}$$

and

$$W_2 = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.1 & 0.9 \\ 0 & 0 & 0.1 & 0.9 \end{pmatrix}.$$

If the two chains start from state 3 or 4, then it is possible to distinguish between them. If, however, 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 $c_r > 0$ depending on the transition matrix of the source such that

$$\limsup_{m\to\infty} \frac{1}{m} \log \bar{D}_m \le -c_r.$$

Remark: The exact form of the constant c_r is given in (17) in the proof of the theorem. In principle, c_r can be explicitly evaluated from the N-block transition probabilities of the process $\{X_{jN}^{(j+1)N-1}\}_{j=0}^{\infty}$, although for larger source alphabets $\mathcal{A}$ and block sizes N the numerical computation may prove prohibitively complex.

Proof of Theorem 2: First notice that $U_0^{\infty} = \{U_{2kN}^{2(k+1)N-1}\}_{k=0}^{\infty}$ is a block-$2N$ stationary 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 Lemma 1 to estimate 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 $\tau = t, t \in \{0, \ldots, N-1\}$ and $U_0^{2N-1} = w \in \mathcal{A}^{2N}$, let $\mathcal{I}_{t,w} \subset \mathcal{A}^{2N}$ denote the set of states reachable from w by the Markov chain. Moreover, for any state w of the chain, let $Q_{t,w} = \{q_t(v|u)\}, u, v \in \mathcal{I}_{t,w}$ denote the transition matrix corresponding to the states in $\mathcal{I}_{t,w}$. That is,

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

To simplify further notation, we extend the above definition to any integer t. Note that $Q_{t,w}$ is a submatrix 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. As $Q_{t,w}$ is defined by the corresponding index set $\mathcal{I}_{t,w}$ and the values $q_t(u,v), u, v \in \mathcal{I}_{t,w}$, we say that $Q_{t,w} = Q_{t',w'}$ if equality holds for both the index sets and the matrix entries, that is $\mathcal{I}_{t,w} = \mathcal{I}_{t',w'}$ and $q_t(v|u) = q_{t'}(v|u)$ for all $u, v \in \mathcal{I}_{t,w}$. Since $X_{-\infty}^{\infty}$ is block-N stationary, $Q_{kN+t,w} = Q_{t,w}$ for any integer $k, t \in \{0, \ldots, N-1\}$ and $w \in \mathcal{A}^{2N}$.

In the proof we will try to estimate which $Q_{t,w}$ is the generator matrix of an observed sequence U_0^m. Note that the block stationarity of U_0^{∞} implies that the starting state of the chain is almost surely nontransient (as the stationary probability of any transient state is 0),[1] that is, if the transition matrix is $Q_{t,w}$ and $U_0^{2N-1} = w$, then, with probability 1, w is a nontransient state of the Markov chain. Clearly, the definition of $\mathcal{I}_{t,w}$ implies that if w is a nontransient state, then $\mathcal{I}_{t,w}$ is an irreducible set of states and $Q_{t,w}$ is irreducible. Therefore, in what follows we only consider those $Q_{t,w}$ matrices that are irreducible.

Obviously, if the $Q_{t,w}$ are not all different, it is not possible to determine the parameters w and t of the real transition matrix exactly (we cannot distinguish between two Markov chains with the same transition matrix). Therefore, for any t let $g_w(t)$ denote the smallest number in $\{0, \ldots, N-1\}$ such that $Q_{t,w} = Q_{g_w(t),w}$, and let N_w^* be the number of different irreducible transition matrices $Q_{t,w}$. (Clearly, N_w^* is almost surely at least one, and it is bounded from above by the number $\hat{N}_w$ of different—not necessarily irreducible—transition matrices $Q_{t,w}$, and it is easy to show that $\hat{N}_w = \max_{0 \le t < N} g_w(t) + 1$, and $Q_{0,w}, \ldots, Q_{\hat{N}_w - 1,w}$ are different.)

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

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

[1]A state of a Markov chain is called nontransient if the chain, started from that given state, returns to that state infinitely many times with probability one, and transient otherwise.

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

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

Therefore,

$$\begin{aligned} &\limsup_{m\to\infty} \frac{1}{m} \log \bar{D}_m \le \\ &\max_{w: P_{U_0^{2N-1}}(w) > 0} \limsup_{m\to\infty} \frac{1}{m} \log H\left(g_w(\tau)|U_{2N}^{m-1}, U_0^{2N-1} = w\right). \end{aligned} \quad (8)$$

Next we bound the conditional entropies

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

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

$$\begin{aligned} &\Pr(g_w(\tau) = \tau_{m,w}|U_0^{2N-1} = w) \\ &\qquad \ge \Pr(g_w(\tau) = f(U_0^{m-1})|U_0^{2N-1} = w) \end{aligned}$$

for any function $f : \mathcal{A}^m \to \{0, \ldots, \hat{N}_w - 1\}$, and let

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

Note that such an optimal estimate always exists, and $Q_{\tau_{m,w},w}$ is almost surely an irreducible matrix. Moreover, the latter implies that if $N_w^* = 1$, then $\tau = \tau_{m,w}$ for all m with probability 1.

Since $\tau_{m,w}$ is a function of U_0^m

$$H\left(g_w(\tau)|U_{2N}^{m-1}, U_0^{2N-1} = w\right) \le H\left(g_w(\tau)|\tau_{m,w}, U_0^{2N-1} = w\right) \quad (9)$$

(for properties of the entropy function see, e.g., [5]). Therefore, if $N_w^* = 1$, then $H(g_w(\tau)|\tau_{m,w}, U_0^{2N-1} = w) = 0$, and so

$$\lim_{m\to\infty} \frac{1}{m} \log H(g_w(\tau)|U_{2N}^{m-1}, U_0^{2N-1} = w) \le -c_w \quad (10)$$

with $c_w = \infty$.

Otherwise, if $N_w^* > 1$, we bound the right-hand side of (9) using Fano's inequality as

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

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

$$\begin{aligned}\limsup_{m\to\infty} \frac{1}{m}\log H(g_w(\tau)|\tau_{m,w}, U_0^{2N-1} = w) \\ \le \limsup_{m\to\infty}\frac{1}{m}\log(2\max\{p_{m,w}\log(N_w^*-1), h_b(p_{m,w})\}) \\ \le \max\left\{\limsup_{m\to\infty}\frac{1}{m}\log p_{m,w}, \limsup_{m\to\infty}\frac{1}{m}\log h_b(p_{m,w})\right\}. \quad (11)\end{aligned}$$

Next we use Lemma 1 to bound (11). 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. As we have mentioned before, if $Q_{t,w}$ is not irreducible, then w is a transient state of the Markov chain, and so its stationary probability is 0. If $Q_{t,w}$ is irreducible, the probability that the corresponding Markov chain does not reach a given state within k steps converges to 0 exponentially fast in k. To determine the exponent, consider the matrices $Q_{t,w,-v}, v \in \mathcal{I}_{t,w}$, obtained from $Q_{t,w}$ by setting its vth column to zero. Since this modification results in throwing away the probability that the chain would reach state v in each step, it is easy to see that the probability of not reaching state v within k steps is given by

$$\hat{p}^{(k)}_{Q_t,w,w,v} = e_w^T Q^k_{t,w,-v} e$$

where $e_w = (0,\dots,0,1,0\dots,0)$ is a unit vector with a 1 at position w, and e is a column vector whose entries are all 1's. Now the irreducibility of $Q_{t,w}$ implies that the probability that the chain, started from any state $w' \in \mathcal{I}_{t,w}$, reaches v converges to one as the number of steps k increases. Thus, $e_{w'}^T Q^k_{t,w,-v} e \to 0$ as $k \to \infty$ for all $w' \in \mathcal{I}_{t,w}$, which shows that $Q^k_{t,w,-v}$ converges to the zero matrix as k increases. This implies that the spectral radius $\rho(Q_{t,w,-v}) < 1$ [14, Theorem 5.6.12]. Then, similarly to (4), we have

$$\limsup_{k\to\infty}\frac{1}{k}\log \hat{p}^{(k)}_{Q_t,w,w,v} \le \log\rho(Q_{t,w,-v}) < 0.$$

Therefore, if $Q_{t,w}$ is irreducible, then 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

$$\begin{aligned}\limsup_{k\to\infty}\frac{1}{k}\log\Pr\left(\hat{\mathcal{I}}_k \ne \mathcal{I}_{t,w}|\tau = t, U_0^{2N-1} = w\right) \\ \le \max_{v\in\mathcal{I}_{t,w}\setminus\{w\}}\log\rho(Q_{t,w,-v}) < 0\end{aligned}$$

by the union bound, since reaching all states requires reaching each state $v \in \mathcal{I}_{t,w}\setminus\{w\}$ individually. This implies that for any w such that $\Pr(U_0^{2N-1} = w) > 0$

$$\limsup_{m\to\infty}\frac{1}{m}\log\Pr\left(\hat{\mathcal{I}}_{\lfloor\frac{m}{2N}\rfloor} \ne \mathcal{I}_{g_w(\tau),w}|U_0^{2N-1} = w\right) \le -c_{w,1} \quad (12)$$

where

$$c_{w,1} = -\frac{1}{2N}\max_{t:Q_{t,w}\text{ is irreducible}}\max_{v\in\mathcal{I}_{t,w}\setminus\{w\}}\log\rho(Q_{t,w,-v}) > 0. \quad (13)$$

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

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

denote the set of indices of the irreducible Markov chains with state space $\mathcal{I}$ (from among Markov chains with the same transition matrix, we pick the one with the smallest index). Now $g_w(\tau)$ can be estimated by first estimating $\mathcal{I}_{g_w(\tau),w}$ by $\hat{\mathcal{I}}_k, k = \lfloor m/2N\rfloor$, based on U_0^{m-1}, and then estimating $g_w(\tau)$ by an optimal classifier for the problem of deciding which $Q_{i,w}, i \in \mathcal{Q}_w(\hat{\mathcal{I}}_k)$, has 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 latter optimal classifier given $U_0^{2N-1} = w$. Then for any w with $\Pr(U_0^{2N-1} = w) > 0$

$$p_{m,w} \le \Pr\left(\hat{\mathcal{I}}_k \ne \mathcal{I}_{g_w(\tau),w}|U_0^{2N-1} = w\right) + p'_{m,w}. \quad (14)$$

Here the first term on the right-hand side is asymptotically bounded by (12). To bound the second term, define

$$c_{w,2} = -\max_{i,j\in\mathcal{Q}_w(\mathcal{I}_{g_w(\tau),w}),i\ne j}\min_{0\le\lambda\le1}\log\rho(Q_{i,j,w,\lambda}) \quad (15)$$

where the matrices $Q_{i,j,w,\lambda}$, whose entries are also indexed by $\mathcal{I}^2_{g_w(\tau),w}$, are given for any $i,j \in \mathcal{Q}_w(\mathcal{I}_{g_w(\tau),w})$ and $0 \le \lambda \le 1$ by

$$Q_{i,j,w,\lambda}(v|u) = Q_{i,w}(v|u)^\lambda Q_{j,w}(v|u)^{1-\lambda}, \qquad u, v \in \mathcal{I}_{g_w(\tau),w}.$$

Then, as the matrices $Q_{i,w}, i \in \mathcal{Q}_w(\mathcal{I}_{g_w(\tau),w})$ are different and irreducible, from Lemma 1 we have $c_{w,2} > 0$ and

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

Combining this inequality with (12) and (14) we obtain

$$\limsup_{m\to\infty}\frac{1}{m}\log p_{m,w} \le -c_w \quad (16)$$

for the positive number $c_w = \min\{c_{w,1}, c_{w,2}\}$. In particular, $\lim_{m\to\infty} p_{m,w} = 0$. Therefore, as L'Hospital's rule implies

$$\lim_{p\to0} p\log(1/p)/h_b(p) = 1$$

we have

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

where the second equality holds because $\lim_{x\to 0^+} \frac{\log\log(1/x)}{\log x} = 0$. Thus, the two terms in the argument of the maximum in (11) are equal and converge to zero exponentially fast by (16). Combining this fact with inequalities (8)–(10) proves the theorem with constant

$$c_r = \min_{w: P_{U_0^{2N-1}}(w)>0} c_w > 0 \tag{17}$$

where $c_w = \infty$ if $N_w^* = 1$, and $c_w = \min\{c_{w,1}, c_{w,2}\}$ if $N_w^* > 1$, where $c_{w,1}$ and $c_{w,2}$ are defined in (13) and (15), respectively. □

V. 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 $\ell_n^{(m)} : \mathcal{A}^n \to \{0,1\}^*$ denote the code lengths of a universal code $\{f_n\}$ for mth-order Markov sources satisfying

$$\frac{1}{n} \sup_{P_{Y_0^{n-1}}} \sup_{z_0^{n-1}\in\mathcal{A}^n} \left[\ell_n^{(m)}\left(z_0^{n-1}\right) + \log P_{Y_0^{n-1}}\left(z_0^{n-1}\right)\right] \le c_n^{(m)} \tag{18}$$

for some $c_n^{(m)} \to 0$ as $n \to \infty$, where the first supremum is taken over all n-fold marginal distributions of mth-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 mth-order Markov source. For example, there exist universal arithmetic codes for mth-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

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

The next result establishes an upper bound on this quantity under the above general conditions.

Theorem 3: If the code length function $\ell_n^{(m)}$ satisfies (18) then for $n \ge m \ge 2N$ the coding redundancy $R_{n,m}$ for the block stationary block Markov source X_0^∞ can be bounded as

$$R_{n,m} \le \frac{1}{n}\log N + 2^{-mc_r+o(m)} + c_n^{(m)} \tag{19}$$

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,

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

ii) It is easy to see that to minimize the bound (19), 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 $m = (\log n - \log\log n - \log|\mathcal{A}|)$, yielding a redundancy bound of order $(n/(A\log n))^{-\frac{c_r}{c_r+\log|\mathcal{A}|}}$. 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 mth-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 (19) is usually the last one.

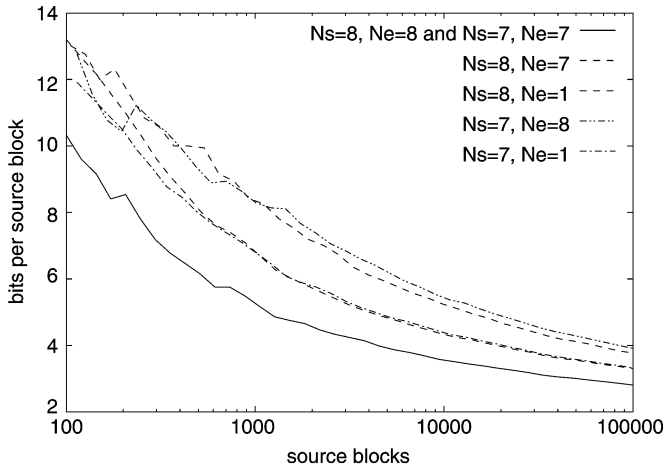

Fig. 1. Compression performance of `bzip2` on "book1" for different source and encoder block sizes.

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).[2] 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.

Proof of Theorem 3: From (18) it follows that for any mth-order Markov source Y_0^∞ and x_0^{n-1}

$$\log \frac{P_{Y_0^{n-1}}\left(x_0^{n-1}\right)}{P_{\ell_n^{(m)}}\left(x_0^{n-1}\right)} \le nc_n^{(m)} \tag{20}$$

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

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

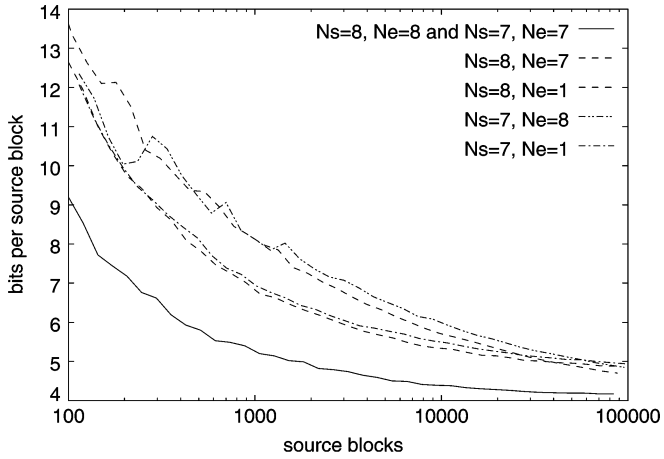

Fig. 2. Compression performance of `bzip2` on the first-order Markov source generated from "book1" for different source and encoder block sizes.

Let $\hat{Y}_0^\infty$ denote the stationary mth-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

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

where the inequality holds by (20). Now, the first term can be easily bounded following the proof of Theorem 1 as

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

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

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

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

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

VI. 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. A. Nagy and T. Linder, "Experimental study of a binary block sorting compression scheme," in *Proc. Data Compression Conf.*, Snowbird, UT, Apr. 2003, p. 439.
[2] M. Burrows and D. J. Wheeler, A block-sorting lossless data compression algorithm. DSRC Res. Rep. 124, Palo Alto, CA, May 1994.
[3] I. Witten and T. Bell, The Calgary Text Compression Corpus [Online]. Available: 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, no. 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. Inf. Theory*, vol. IT-31, no. 3, pp. 360–365, May 1985.
[9] V. Anantharam, "A large deviations approach to error exponents in source coding and hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 36, no. 4, pp. 938–943, Jul. 1990.
[10] N. Merhav and J. Ziv, "A Bayesian approach for classification of Markov sources," *IEEE Trans. Inf. Theory*, vol. 37, no. 4, pp. 1067–1071, Jul. 1991.
[11] I. Csiszár, T. M. Cover, and B.-S. Choi, "Conditional limit theorems under Markov conditioning," *IEEE Trans. Inf. Theory*, vol. IT-33, no. 6, pp. 788–801, Nov. 1987.
[12] A. György, D. A. Nagy, and T. Linder, "Convergence rates in higher-order Markov modeling of block-Markov sources," in *Proc. Canadian Workshop on Information Theory* , Montreal, QC, Canada, Jun. 2005, pp. 111–114.
[13] D. A. Nagy, "Lossless compression and alphabet size," Ph.D. dissertation, Queen's Univ., Kingston, ON, Canada, 2006.
[14] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.
[15] P. Lancaster, *Theory of Matrices*. New York: Academic, 1969.