Convergence Rates in Higher Order Markov Modeling of Block-Markov Sources

András György
Informatics Laboratory
Computer and Automation Research Institute
Hungarian Academy of Sciences
Lágymányosi u. 11, 1111 Budapest, Hungary
Email: gya@szit.bme.hu

Daniel A. Nagy
Dept. of Mathematics and Statistics
Queen's University
Kingston, Ontario
Canada, K7L 3N6
nagydani@mast.queensu.ca

Tamás Linder
Dept. of Mathematics and Statistics
Queen's University
Kingston, Ontario
Canada, K7L 3N6
linder@mast.queensu.ca

I. Introduction

In practice, lossless compression is often applied to data that is composed of fixed-length blocks of elementary symbols. For example, sequences of bytes (consisting of 8 bits) or sequences of 32-bit words (consisting of 4 bytes). As suggested by experimental data presented in [1], universal compression algorithms operating on elementary symbols (such as bits) achieve good compression results over a large family of sources, while those geared towards one particular block size are restricted to that and its multiples.

In this paper, we investigate the relative entropy rate of a higher-order Markov model and a first-order block-Markov source. In particular, we show that as the order of the Markov model increases, this relative entropy rate converges to zero exponentially fast. Since the redundancy of a variable-length lossless code matched to a model is essentially the relative entropy between the source model distribution and the actual distribution of the source, our result suggests that modeling data on the elementary symbol level not taking blocks into account is acceptably low in a Markovian setting.

II. Preliminaries

A sequence of random variables $\{X_n\}_{n=0}^{\infty}$ taking values in a finite alphabet A is said to be a block-N-Markov source if the N-blocks $X_0^{N-1}, X_N^{2N-1}, \ldots$ form a Markov chain. The sequence $\{Y_n\}_{n=0}^{\infty}$ taking values in A is called an mth order Markov source if for any $M \geq m$, the conditional distribution of Y_M given Y_0^{M-1} is the same as the conditional distribution of Y_M given Y_{M-m}^{M-1} and these conditional distributions do not depend on M.

We assume that the initial segments X_0^{N-1} and Y_0^{m-1} are both drawn from the stationary distributions of the respective processes (which we assume to exist), so that Y_0^{∞} is a stationary process and X_0^{∞} is block-N stationary process.

The divergence rate (relative entropy rate) between the two sources is defined as usual:

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

assuming the limit exists. In our case, both sources are stationary block-Markov processes with a block-size that is a multiple of both N and m, for which the limit always exists.

To determine the rate loss (redundancy) of lossless coding which results from using the best mth order Markov model on symbols of A in approximating the actual block-N-Markov source (with symbols in A^N), one has to find

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

In [2], the best approximating mth order Markov-model Y has been found for a given block-N-Markov source X and its divergence rate has been calculated for sufficiently large m:

Proposition 1 ([2]): Given a block-N Markov source X_0^{∞}, the relative entropy rate $\bar{D}(X_0^{\infty} \,\|\, Y_0^{\infty})$ is minimized over all mth order Markov sources Y_0^{∞} if and only if $P_{Y_0^m} = P_{U_0^m}$, where the random vector $U_0^m = U_0, \ldots, U_m$ is defined by $U_j = X_{j-m+\tau}, \quad j = 0, \ldots, m$ where τ is a random variable that is uniformly distributed on $\{0, 1, \ldots, N-1\}$ and is independent of $\{X_n\}$. The resulting 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}.

A simple consequence of this result is that as m increases, the minimum relative entropy rate converges to zero, i.e., $\lim_{m\to\infty} \bar{D}_m = 0$.

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC), the NATO Science Fellowship, and by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

III. Results

The following new result shows how fast $\bar{D}_m$ converges to zero.

Theorem 1: For every irreducible block-stationary block-Markov source X_0^∞ there is a constant $c > 0$ depending on the transition matrix of the source such that

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

Note that this implies $\bar{D}_m = o(2^{-(c-\epsilon)m})$ for any $\epsilon > 0$.

The proof of Theorem 1 hinges on the observation that for any given value of τ, U_0^∞ is a block-$2N$-Markov source, thus the problem of estimating τ based on the sequence U_0^{m-1} can be reformulated as a Markov-source classification problem [3].

Proof sketch: For $t \equiv \tau (\bmod N)$, $t \in \mathbb{N}$, let $Q_t = \{q_t(u|v)\}, u, v \in A^{2N}$ denote the probability transition matrix of U_{2N}^{4N-1} given U_0^{2N-1} and $\tau = t \bmod N$.

Note that $Q_{kN+t} = Q_t$ for any integer k and $t \in \{0, \dots, N-1\}$, since $X_{-\infty}^\infty$ is block-N-stationary. Let $N^* > 0$ be the smallest positive integer such that $Q_t = Q_{t+N^*}$ for some $t \in \{0, \dots, N-1\}$. Note that $N^* \le N$ as $Q_t = Q_{t+N}$ for all t. Since X_0^∞ is block-N-stationary, it can be easily seen that $Q_s = Q_{s+N^*}$ for all s. Since $Q_s = Q_{s+N}$ for all s, this, and the minimal property of N^* imply that $Q_t = Q_{t'}$ if and only if N^* divides $t - t'$. In particular, it follows that N^* divides N.

Thus, $Q_0, \dots, Q_{N^*-1}$ are different. Moreover, since $X_{-\infty}^\infty$ is irreducible and $Q_t = Q_{t+N^*}$ for all t, $X_{-\infty}^\infty$ is also block-N^*-stationary. Therefore, $\lfloor \frac{\tau}{N^*} \rfloor$ is independent of U_0^m for any m. That is, if $\tau = \alpha N^* + \tau^*$ with $\tau^* \in \{0, \dots, N^*-1\}$, then α is independent of (τ^*, U_0^∞) and is uniformly distributed over $\{0, \dots, N/N^* - 1\}$, and τ^* is uniformly distributed over $\{0, \dots, N^* - 1\}$. Therefore, $H(\tau|U_0^m) = \log N - \log N^* + H(\tau^*|U_0^m)$ for all m, and

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

If $N^* = 1$, then $X_{-\infty}^\infty$ is an Nth order Markov source. In this case, $\bar{D}_m = 0$ for all $m \ge N$, and the theorem holds with $c = \infty$. Otherwise, if $N^* > 1$, let $\tau_m = \tau_m(U_0^{m-1})$ be an optimal estimate of τ^* based on U_0^{m-1} in the sense that $\Pr(\tau^* = \tau_m) \ge \Pr(\tau^* = f(U_0^{m-1}))$ for any function $f : A^m \to \{0, \dots, N^* - 1\}$, and let $p_m = \Pr(\tau^* \ne \tau_m)$. Then it follows from (1) that

$$\begin{aligned}\bar{D}_m &= H(\tau^*|U_0^{m-1}) - H(\tau^*|U_0^m) \\ &\le H(\tau^*|U_0^{m-1}) \le H(\tau^*|\tau_m). \quad (2)\end{aligned}$$

By Fano's inequality $H(\tau^*|\tau_m) \le \log(N^* - 1)p_m + h_b(p_m)$ where $h_b(p) = -p\log p - (1-p)\log(1-p)$ for $0 \le p \le 1$. From here, it is not hard to see that

$$\begin{aligned}&\limsup_{m\to\infty} \frac{1}{m} \log H(\tau^*|\tau_m) \quad (3) \\ &\le \max\left\{ \limsup_{m\to\infty} \frac{1}{m} \log p_m,\ \limsup_{m\to\infty} \frac{1}{m} \log h_b(p_m) \right\}.\end{aligned}$$

Now, using the techniques of [3], it can be shown that one can estimate, with probability that is exponentially small in m, which of the $Q_0, \dots, Q_{N^*-1}$ generated U_0^m (recall that $Q_0, \dots, Q_{N^*-1}$ are different). This means that there is an estimate of τ which has error probability exponentially small in m. In particular, this implies that

$$\limsup_{m\to\infty} m^{-1} \log p_m \le -c$$

for some $c > 0$.

Now we bound the second term in the maximum in (3).

$$h_b(p_m) \le p_m(1 - p_m - \log p_m) \quad (4)$$

where the inequality holds since $\log x \le (x-1)/\ln 2$ for all $x > 0$. Then, from (4) we obtain

$$\limsup_{m\to\infty} m^{-1} h_b(p_m) \le -c. \quad (5)$$

Combining inequalities (2)-(5) proves the theorem. □

IV. 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 where one can take advantage of the computational benefits of operating on the smallest possible alphabet.

Note that we have made the simplifying assumption that the bit-level algorithm is optimal for the mth order model. In practice, however, one may only hope that the bit-level code is only asymptotically optimal, i.e., universal in the class of finite order Markov sources. In this case, further investigation is needed to relate the coding redundancy to the relative entropy rate we have studied.

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] D. Nagy and T. Linder, "Markov Modeling of Block-Markov Sources," in *proceedings of Queen's Biennial Symposium on Communications, QBS'04,* Kingston, ON, Canada, p. 118–120, June 2004.

[3] 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.