Higher Order Markov Modeling of Block-Markov Sources[1]

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

Tamás Linder
Dept. of Mathematics & Statistics
Dept. of Elec. & Comp. Engineering
Queen's University,
Kingston, Ontario K7L 3N6
email: linder@mast.queensu.ca

Abstract — **Industry-standard lossless compression algorithms (such as LZW [2]) 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 [3] while algorithms working on a bit level perform reasonably on byte-based data. In present paper, we offer an information-theoretic explanation to these experimental results by assessing the redundancy (which is approximated by the divergence rate of two distributions) of a bit-based model when applied to a byte-based source. More specifically, we apply a higher order Markov model to a block-Markov source, as higher-order Markov sources of bytes are also block Markov sources with a block-size that is a multiple of 8 bits. We show that the divergence rate between a block-Markov source and the best-matching higher order Markov model for that source is small enough for practical coding purposes. Namely, the divergence rate between a an order m Markov model and a block-N Markov source can be as small as the mutual information between the $(m+1)$st symbol and its randomly selected position within the N-long block given the past m symbols. This quantity is bounded from above by $\log N$ and converges to zero as m increases without bound.**

I. Motivation

Lossless data compression is the science of representing digital information 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 does not hold (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 byte-aligned sources seems acceptably low [3]. In present paper, we evaluate this penalty in a somewhat more general 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 [1]. The excess of code rate over the entropy rate is called the (rate) redundancy of the code. This is the quantity that needs to be minimized when designing a lossless code. For the sake of simplicity, we assume that the code is optimal for the model distribution. In this setting, the relative entropy rate [4] between the model distribution and that of the source approximates the rate redundancy of the code with respect to the source. Thus, for assessing the rate redundancy, it is important to determine the relative entropy rate between different source models. In this work, we analyze the divergence rate between higher order Markov models and block-Markov sources, and show that higher order Markov models can efficiently model block-Markov sources.

II. Preliminaries

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

$$D(Z\|W) = D(P_X\|P_W) = \sum_{z\in\mathcal{Z}} P_Z(z) \log \frac{P_Z(z)}{P_W(z)}$$

where $\log(\cdot)$ denotes base 2 logarithm. Note that $D(P_X\|P_W)$ is nonnegative and equals zero if and only if $P_Z = P_W$ [4].

The relative entropy between pairs of random variables $Z_1^2 = (Z_1, Z_2)$ and $Z_1^2 = (W_1, W_2)$ can be expressed, using the chain rule [4], as

$$D(P_{Z_1^2} \,\|\, P_{W_1^2}) = D(P_{Z_1} \,\|\, P_{W_1}) + D(P_{Z_2|Z_1} \,\|\, P_{W_2|W_1})$$

where

$$\begin{aligned} D(P_{Z_2|Z_1}\|P_{W_2|W_1}) &\triangleq \sum_{z_1} D(P_{Z_2|Z_1}(\cdot|z_1) \,\|\, P_{W_2|W_1}(\cdot|z_1))P_{Z_1}(z_1) \\ &= \sum_{z_1} P_{Z_1}(z_1) \sum_{z_2} P_{Z_2|Z_1}(z_2|z_1) \log \frac{P_{Z_2|Z_1}(z_2|z_1)}{P_{W_2|W_1}(z_2|z_1)}. \end{aligned}$$

For any sequence of random variables $\{X_n\}_{n=0}^{\infty} = X_0, X_1, \ldots, X_n, \ldots$ and for any $i \geq j$, 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

[1]This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

$\{X_n\}_{n=0}^{\infty}$. A similar convention applies to deterministic sequences which are usually denoted by lower case letters.

The 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 for all nonnegative integers i and block of symbols $x_0^{(i+1)N-1} \in A^{(i+1)N}$,

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

A sequence of random variables $\{Y_n\}_{n=0}^{\infty}$ taking values in A is called an mth order Markov source if

$$P_{Y_{i+m}|Y_i^{i+m-1}}(x_m \mid x_0^{m-1}) = P_{Y_{j+m}|Y_j^{j+m-1}}(x_m \mid x_0^{m-1})$$

for all nonnegative integers i and j and $x_0^m \in A^{m+1}$.

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 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 [5].

III. Approximation of block Markov sources

We assume that the block-N-Markov source X_0^{∞} is given and look for the minimum divergence rate between X_0^{∞} and any mth order Markov source Y_0^{∞}. That is, we want to determine

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

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 next proposition expresses the minimum divergence rate in terms of the random variables $U_0^m = U_0, \ldots, U_m$ 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 $\{X_n\}$.

Proposition 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 mth order Markov sources Y_0^{∞} if and only if $P_{Y_0^m} = P_{U_0^m}$. 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_0 given U_0^{m-1}.

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

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

Thus we have 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*

$$\sum_{m=1}^{\infty} \bar{D}_m \leq \log N.$$

In particular

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

Remark 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., [5]). Note, however, that X_0^{∞} is non-stationary, and that the proposition gives an explicit expression for the optimum approximating process and a characterization of the resulting minimum divergence rate $\bar{D}_m$, which can be used to determine the rate at which $\bar{D}_m$ converges to zero.

Proof of Proposition 1 For all $n > m$ we have from the chain rule for the relative entropy

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

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

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

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

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

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

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

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

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

which was to be shown. □

IV. Conclusion

We have demonstrated that block-Markov sources can be encoded with vanishing redundancy using codes that are optimized for higher-order symbol-level Markov models. This partially explains the findings of our past experiments [3] 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.

Some theoretical issues remain open. First, further studies are needed to establish the speed at which the relative entropy rate between a block-Markov source and the best-matching mth order Markov model converges to zero. Also, 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] C. E. Shannon: "A mathematical theory of communication" *Bell Sys. Tech. Journal*, 27: 379-423, 623-656, 1948.

[2] M. Nelson and J. Gailly: "The Data Compression Book, 2nd edition" *M&T books*, New York, 1995

[3] D. Nagy and T. Linder, "Experimental study of a binary block sorting compression scheme," *Proceedings of Data Compression Conference, DCC'03*, (Snowbird, UT, 2003), IEEE Comput. Soc. Press, p. 439.

[4] T. M. Cover and J. A. Thomas: *Elements of Information Theory*. Wiley Series in Telecommunications, New York, 1991

[5] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.