Second-Order Noiseless Source Coding Theorems

I. Kontoyiannis, *Student Member, IEEE*

***Abstract*—Shannon's celebrated source coding theorem can be viewed as a "one-sided law of large numbers." We formulate second-order noiseless source coding theorems for the deviation of the codeword lengths from the entropy. For a class of sources that includes Markov chains we prove a "one-sided central limit theorem" and a law of the iterated logarithm.**

***Index Terms*—Coding variance, convergence rates, source coding theorems.**

I. Introduction

Let $\boldsymbol{X} = \{X_n;\, n \in \mathbb{Z}\}$ be a stationary ergodic source with finite alphabet A, and let $L_n : A^n \to \mathbb{N}$ be an arbitrary sequence of fixed-to-variable codeword length assignments. Let X_1^n denote the block $(X_1, X_2, \cdots, X_n)$ and H be the entropy rate of $\boldsymbol{X}$. From the pointwise converse source coding theorem [2], [6] we know that, eventually, the per-symbol codeword length will exceed H, along almost any source realization

$$\liminf_{n\to\infty} \frac{L_n(X_1^n)}{n} \geq H \quad \text{a.s.} \tag{1}$$

We also know that there exists a sequence $\{L_n^*\}$, such that the above lower bound is actually met with equality

$$\liminf_{n\to\infty} \frac{L_n^*(X_1^n)}{n} = H \quad \text{a.s.} \tag{2}$$

Equation (2) tells us that the average number of bits required to describe X_1^n converges to H, with probability one. In a formal sense, this can be thought of as a "strong law of large numbers" for the codeword lengths, and, similarly, the corresponding asymptotic lower bound (1) as a "one-sided" law of large numbers.

It is then a natural question to ask whether this relationship can be refined to a "one-sided" central limit theorem (CLT) or a law of the iterated logarithm (LIL). In this correspondence we show that when $\boldsymbol{X}$ is a Markov chain, or, more generally, when $\boldsymbol{X}$ satisfies certain mixing conditions, the answer to this question is affirmative.

In the next section we state and discuss our main results, a one-sided CLT and a LIL for the codeword lengths $\{L_n\}$. In Section III we give their proofs, and Section IV discusses their extensions to non-Markov sources. In the Appendix we give the proof of an unpublished result that is used in Section III.

II. Results

Let $\boldsymbol{X} = \{X_n;\, n \in \mathbb{Z}\}$ be a stationary ergodic Markov chain with finite alphabet A, distributed according to the measure P. Let $C_n : A^n \to \{0, 1\}^*$, $n \geq 1$, be an arbitrary sequence of fixed-to-variable length prefix codes (not necessarily mutually compatible), and $L_n : A^n \to \mathbb{N}$, $n \geq 1$, be the associated sequence of length functions. Let $H = E(-\log_2 P(X_1|X_0))$ denote the entropy rate of $\boldsymbol{X}$. (Here and throughout the paper "$\log_2$" denotes the logarithm to base two and "log" denotes the natural logarithm.) For the sake of simplicity we will assume that $\boldsymbol{X}$ is a first-order Markov chain, although this restriction is not necessary and it will be dropped later on (the extension of our results to Markov chains of any finite order, as well as to non-Markov sources under mixing conditions is outlined in Section IV).

We are interested in the asymptotic behavior of the deviation of the codeword lengths $L_n(X_1^n)$ from their "ideal means" $H(X_1^n) = E(-\log_2 P(X_1^n))$; if we let $D_n = [L_n(X_1^n) - H(X_1^n)]$ then the converse coding theorem (1) can be restated as

$$\liminf_{n\to\infty} \frac{D_n}{n} \geq 0 \quad \text{a.s.}$$

Our first result is a CLT refinement to this. If instead of normalizing by n we normalize by $\sqrt{n}$ then the deviation $D_n/\sqrt{n}$ is (asymptotically) bounded below by a sequence of random variables whose distribution converges to a Gaussian.

Theorem 1(CLT): Let $\boldsymbol{X}$ be a stationary ergodic Markov chain with alphabet A, and suppose $\{L_n\}$ is an arbitrary sequence of codeword-length assignments $L_n : A^n \to \mathbb{N}$. Define

$$D_n = L_n(X_1^n) - H(X_1^n).$$

There exists a sequence of random variables Z_n such that

$$\liminf_{n\to\infty} \left[\frac{D_n}{\sqrt{n}} - Z_n\right] \geq 0 \quad \text{a.s.}$$

$$Z_n \xrightarrow{\mathcal{D}} N(0, \sigma^2)$$

and the variance σ^2 is given by the limit

$$\sigma^2 = \lim_{n\to\infty} \frac{1}{n} \mathrm{Var}\,(-\log_2 P(X_1^n)). \tag{3}$$

Our second result is a corresponding law of the iterated logarithm.

Theorem 2 (LIL): Under the assumptions of Theorem 1, and with σ^2 defined as in (3)

$$\text{i)} \qquad \limsup_{n\to\infty} \frac{D_n}{\sqrt{2n \log\log n}} \geq \sigma \quad \text{a.s.}$$

$$\text{ii)} \qquad \liminf_{n\to\infty} \frac{D_n}{\sqrt{2n \log\log n}} \geq -\sigma \quad \text{a.s.}$$

Suppose $\sigma^2 > 0$. Then part i) of Theorem 2 implies that for any sequence $\{L_n\}$ and any constant $K \in (0, \sigma)$, along almost any realization of the source, the codeword lengths $L_n(X_1^n)$ will be greater than

$$H(X_1^n) + K\sqrt{2n \log\log n}$$

infinitely often. The interpretation of this bound is discussed in some more detail in Remark 2 below.

As for the case $\sigma^2 = 0$, a complete characterization is provided by the following theorem. If was first stated in [12], and a proof was supplied in [7].

Theorem 3 [12], [7]: Suppose $\boldsymbol{X}$ is a stationary ergodic Markov chain, and let σ^2 be defined as in Theorem 1. Then $\sigma^2 = 0$ if and only if all the nonzero transition probabilities of the chain are equal to 2^{-H}.

Theorem 3 says that $\sigma^2 = 0$ if and only if, for each n, there are $2^{nH} \leq |A|^n$ sequences of nonzero probability, and they are uniformly distributed. We can encode them in such a way that they all have equal-length descriptions (Shannon code), so that the variance of the codeword lengths is zero. In fact, from the proof of Theorem 1 (see

Manuscript received October 28, 1996; revised January 22, 1997. This work was supported in part by NSF under Grant NCR-9205663, by JSEP under Grant DAAH04-94-G-0058, and by ARPA under Grant J-FBI-94-218-2.

The author is with the Information Systems Laboratory, Electrical Engineering Department, Stanford University, Stanford, CA 94305-4055 USA.

Publisher Item Identifier S 0018-9448(97)04393-9.

(5) and (6)), it is clear that when $\sigma^2 = 0$ a slightly stronger statement than Theorems 1 and 2 can be made

$$\liminf_{n\to\infty} \frac{D_n}{\sqrt{n}} \geq 0 \quad \text{a.s.}$$

In view of the preceding comments we think of σ^2 as a *minimal coding variance*. It is a characteristic quantity of the source which tells us that, when we encode the source in the most efficient way, then the asymptotic variance of the codeword lengths will be equal to σ^2. If we do not use the most efficient code, then the deviation of our codeword lengths from the entropy will asymptotically be bounded below by a Gaussian random variable with variance σ^2.

Remarks:

1) Achievability:

In traditional information theoretic terms, Theorems 1 and 2 could be called "second-order converse source coding theorems." But are they ever satisfied with equality? As will become obvious from the proofs (cf. (6) below), equality in all the "almost sure" statements of Theorems 1 and 2 is achieved by the Shannon code: $L_n(X_1^n) = \lceil -\log_2 P(X_1^n) \rceil$.

2) D_n Versus Pointwise Redundancy:

As remarked earlier, from Part i) of Theorem 2 we can deduce a lower bound to the rate of convergence of L_n/n to H. If σ^2 is positive then for any constant $K \in (0, \sigma)$

$$\frac{L_n(X_1^n)}{n} \geq H + K\frac{\sqrt{2n \log\log n}}{n} \text{ infinitely often} \qquad \text{a.s.} \quad (4)$$

Since $\sqrt{2n \log\log n}$ increases much faster than $\log n$ this seems to disagree with the well-known universal coding results [8], [11], that exhibit universal procedures achieving convergence rates of order $(\log n)/n$.

The reason for this discrepancy is that here we are investigating the asymptotic behavior of the quantity D_n, whereas the two main quantities of interest in the universal coding literature are the *pointwise redundancy* and its expected value. The pointwise redundancy R_n is defined as the difference between the actual codeword length L_n and the ideal Shannon codeword length $L_n^*(X_1^n) = -\log_2 P(X_1^n)$:

$$R_n(X_1^n) = L_n(X_1^n) - (-\log_2 P(X_1^n)),$$

and the rate at which R_n tends to zero tells us at which rate our code approaches the performance of the Shannon code. The quantity D_n, on the other hand, is the deviation of the codeword length L_n from the "ideal mean" $H(X_1^n)$

$$D_n(X_1^n) = L_n(X_1^n) - E(-\log_2 P(X_1^n)).$$

It is, therefore, plausible that the quantities R_n and D_n will decrease at different rates. As for the expected redundancy, ER_n, although it is of course equal to ED_n, the pointwise bound given by (4) does not necessarily imply a corresponding bound for the expectations $ED_n = ER_n$ (it is trivial that the expectations of a sequence of random variables can converge much faster than the individual realizations do).

III. Proofs

In this section we give the proofs of Theorem 1 and 2 for the case of first-order Markov chains. Their extension to the general case is straightforward, as discussed in the next section. The proofs are simple, and they will depend on the following Lemma. It is an unpublished result that appeared in [2], and also, in a more general form, in [1]. It is proved in the Appendix.

Lemma [1], [2]: For any sequence $\{c(n)\}$ of positive constants with $\sum 2^{-c(n)} < \infty$ then

$$L_n(X_1^n) \geq -\log_2 P(X_1^n) - c(n) \text{ eventually} \quad \text{a.s.}$$

An elegant application of this Lemma to string matching was given by Shields in [10], where the Lemma is referred to as "Barron's Lemma."

Proof of Theorem 1: Since the series $\sum 2^{-\epsilon\sqrt{n}}$ is finite for any $\epsilon > 0$, we can apply the Lemma with $c(n) = \epsilon\sqrt{n}$ to deduce that

$$\liminf_{n\to\infty} \frac{1}{\sqrt{n}}[L_n(X_1^n) + \log_2 P(X_1^n)] \geq -\epsilon \quad \text{a.s.}$$

and since $\epsilon > 0$ is arbitrary

$$\liminf_{n\to\infty} \frac{1}{\sqrt{n}}[L_n(X_1^n) + \log_2 P(X_1^n)] \geq 0 \quad \text{a.s.}$$

or, equivalently,

$$\liminf_{n\to\infty} \left[\frac{D_n}{\sqrt{n}} - Z_n\right] \geq 0 \quad \text{a.s.} \qquad (5)$$

where

$$Z_n = \frac{-\log_2 P(X_1^n) - H(X_1^n)}{\sqrt{n}}. \qquad (6)$$

Let $S_n = -\log_2 P(X_1^n) - H(X_1^n)$ and consider the Markov chain

$$\tilde{\boldsymbol{X}} = \{\tilde{X}_n = (X_n, X_{n+1});\, n \in \mathbb{Z}\}$$

with alphabet

$$B = \{(i, j) \in A \times A: P(X_{k+1} = j | X_k = i) > 0\}.$$

Using the Markovity of $\boldsymbol{X}$, we can expand S_n as

$$\begin{aligned} S_n &= \sum_{i=1}^{n-1}(-\log_2 P(X_{i+1}|X_i) - H) + (-\log_2 P(X_1) - H(X_1)) \\ &= \sum_{i=1}^{n-1}(f(\tilde{X}_i) - Ef(\tilde{X}_i)) + (-\log_2 P(X_1) - H(X_1)) \qquad (7) \end{aligned}$$

where $f: B \to \mathbb{R}$ is the map

$$(i, j) \mapsto -\log_2 P(X_{n+1} = j | X_n = i).$$

Therefore, the random variables S_n behave (up to a bounded term) like the partial sums of a centered, bounded function of a Markov chain. Since $\boldsymbol{X}$ is stationary ergodic so is $\tilde{\boldsymbol{X}}$, and since B is finite $\tilde{\boldsymbol{X}}$ is irreducible and aperiodic. By the central limit theorem for functions of Markov chains (see [4], for example) the limit

$$\tau^2 = \lim_{n\to\infty} \frac{1}{n} \mathrm{Var}\,(-\log_2 P(X_1^n|X_0))$$

exists and is finite, and since

$$-\log_2 P(X_1^n|X_0) = -\log_2 P(X_1^n) + \log_2[P(X_1)/P(X_1|X_0)]$$

it is easy to show that $\tau^2 = \sigma^2$. Moreover, the first term in (7) normalized by $\sqrt{n}$ converges in distribution to a $N(0, \sigma^2)$ random variable. Since the second term in (7) is bounded (with probability one) we conclude that Z_n converges in distribution to a $N(0, \sigma^2)$ random variable, and this completes the proof. □

Proof of Theorem 2: We proceed in a similar fashion to the proof of Theorem 1. Since the series $\sum 2^{-\epsilon\sqrt{2n \log\log n}}$ is finite for any $\epsilon > 0$, we can apply the Lemma with $c(n) = \epsilon\sqrt{2n \log\log n}$ to deduce that

$$\frac{D_n}{\sqrt{2n \log\log n}} \geq \frac{S_n}{\sqrt{2n \log\log n}} + \epsilon \text{ eventually} \quad \text{a.s.}$$

Taking the $\limsup$ (respectively, $\liminf$) of both sides and letting ϵ decrease to zero yields

$$\text{a)} \qquad \limsup_{n\to\infty} \frac{D_n}{\sqrt{2n \log\log n}} \geq \limsup_{n\to\infty} \frac{S_n}{\sqrt{2n \log\log n}} \quad \text{a.s.}$$

and

$$\text{b)} \qquad \liminf_{n\to\infty} \frac{D_n}{\sqrt{2n \log\log n}} \geq \liminf_{n\to\infty} \frac{S_n}{\sqrt{2n \log\log n}} \quad \text{a.s.}$$

Now arguing as in the proof of Theorem 1, we can divide (7) by $\sqrt{2n \log\log n}$ and then apply the law of the iterated logarithm for functions of Markov chains [4] to get

$$\text{a}') \qquad \limsup_{n\to\infty} \frac{S_n}{\sqrt{2n \log\log n}} = \sigma \quad \text{a.s.}$$

and

$$\text{b}') \qquad \liminf_{n\to\infty} \frac{S_n}{\sqrt{2n \log\log n}} = -\sigma \quad \text{a.s.}$$

Combining a) with a$'$) and b) with b$'$) yields i) and ii), respectively. □

IV. Extensions

The proofs of Theorems 1 and 2 depend on two simple results: The Lemma of the previous section, and the fact that the random walk $S_n = -\log_2 P(X_1^n) - H(X_1^n)$ satisfies a CLT and an LIL. Since the Lemma is true for any source, the only place in the proof where the Markovian assumption is used is to obtain the asymptotic properties of S_n.

Higher Order Markov Chains: If $\boldsymbol{X}$ is Markov of general order $m \geq 1$, we can simply modify our proofs by looking at the chain

$$\tilde{\boldsymbol{X}} = \{\tilde{X}_n = (X_n, X_{n+1}, \cdots, X_{n+m});\, n \in \mathbb{Z}\}$$

and expanding S_n as the sum of the logarithms of mth-order conditional probabilities to get the required CLT and LIL for S_n. All other parts of the proofs remain the same.

Non-Markov Sources: The question of the exact description of the asymptotics of S_n was raised by Kolmogorov in the early 1950's, and was later studied in detail by Yushkevich [12], Ibragimov [5], and Philipp and Stout [9, ch. 9], who obtained an almost sure invariance principle for S_n under certain mixing conditions described below.

For $-\infty \leq i \leq j \leq \infty$ let $\mathcal{B}_i^j$ denote the σ-field generated by the random variables X_i^j, and for $d \geq 1$ define the mixing coefficients

$$\gamma(d) = \max_{s\in S} E\Big|\log_2 P(X_0 = s|X_{-\infty}^{-1}) - \log_2 P(X_0 = x|X_{-d}^{-1})\Big|$$
$$\alpha(d) = \sup\{|P(B \cap A) - P(B)P(A)|;\, A \in \mathcal{B}_{-\infty}^0,\, B \in \mathcal{B}_d^\infty\}.$$

The coefficients $\alpha(d)$ are called the *strong mixing* coefficients of $\boldsymbol{X}$, and the coefficients $\gamma(d)$ were introduced by Ibragimov in [5]. (See [3] for the standard properties of $\alpha(d)$.)

From [9, Theorem 9.1] it follows immediately that if $\boldsymbol{X}$ is a stationary process such that $\alpha(d) = O(d^{-336})$ and $\gamma(d) = O(d^{-48})$, then S_n satisfies a CLT and LIL, with asymptotic variance

$$\sigma^2 = \mathrm{Var}\,(-\log_2 P(X_0|X_{-\infty}^{-1})) + 2\sum_{k=1}^{\infty} \mathrm{Cov}(-\log_2 P(X_0|X_{-\infty}^{-1}), -\log_2 P(X_k|X_{-\infty}^{k-1})). \quad (8)$$

Therefore, we get the following corollary:

Corollary: Theorems 1 and 2 remain valid if the Markovian assumption for $\boldsymbol{X}$ is replaced by the assumptions that $\boldsymbol{X}$ is stationary,

$$\alpha(d) = O(d^{-336})$$
$$\gamma(d) = O(d^{-48})$$

and the expression for the variance σ^2 in (3) is replaced by (8).

Observe that if $\boldsymbol{X}$ is an irreducible aperiodic Markov chain of order $m \geq 1$, then $\gamma(d) = 0$ for all $d \geq m$, the $\alpha(d)$ decay to zero exponentially fast, and the expressions in (3) and (8) coincide, so that the above Corollary is a genuine generalization of Theorems 1 and 2. In the stationary case, the mixing conditions in the Corollary are satisfied by a rather large class of non-Markov processes. Although in practice they may be hard to verify, they require only polynomial decay of the coefficients $\alpha(d)$ and $\gamma(d)$.

Appendix
Proof of the Lemma

We expand the probability

$$\begin{aligned} P\{L_n(X_1^n) < -\log_2 P(X_1^n) - c(n)\} &= P\{x_1^n \in A^n : P(x_1^n) > 2^{-L_n(x_1^n)-c(n)}\} \\ &= \sum_{x_1^n : P(x_1^n) < 2^{-L_n(x_1^n)-c(n)}} P(x_1^n) \\ &\leq \sum_{x_1^n : P(x_1^n) < 2^{-L_n(x_1^n)-c(n)}} 2^{-L_n(x_1^n)-c(n)} \\ &\leq 2^{-c(n)} \sum_{x_1^n \in A^n} 2^{-L_n(x_1^n)} \\ &\leq 2^{-c(n)} \end{aligned}$$

where the last inequality is just Kraft's inequality. Since we assume $\sum 2^{-c(n)} < \infty$, the result follows by the Borel–Cantelli Lemma. □

Acknowledgment

The author wishes to thank Prof. T. Cover for his helpful comments.

References

[1] P. H. Algoet, "Log-optimal investment," Ph.D. dissertation, Dept. Elec. Eng., Stanford Univ., Stanford, CA, 1985.

[2] A. R. Barron, "Logically smooth density estimation," Ph.D. dissertation, Dept. Elec. Eng., Stanford Univ., Stanford, CA, 1985.

[3] B. C. Bradley, "Basic properties of strong mixing conditions," in *Dependence in Probability and Statistics,* E. Wileln and M. S. Taqqu, Eds. Boston, MA: Birkhäuser, 1986, pp. 165–192.

[4] K. L. Chung, *Markov Chains with Stationary Transition Probabilities.* New York: Springer-Verlag, 1967.

[5] I. A. Ibragimov, "Some limit theorems for stationary processes," *Theory Probab. Appl.*, vol. 7, pp. 349–382, 1962.

[6] J. C. Kieffer, "Sample converses in source coding theory," *IEEE Trans. Inform. Theory*, vol. 37, pp. 263–268, Mar. 1991.

[7] I. Kontoyiannis, "Asymptotic recurrence and waiting times for stationary processes," NSF Tech. Rep. 92, Dept. Statist., Stanford Univ., Stanford, CA, submitted for publication, July 1996.

[8] R. E. Krichevsky and V. K. Trofimov, "The performance of universal coding," *IEEE Trans. Inform. Theory,* vol. IT-27, pp. 199–207, Mar. 1981.

[9] W. Phillipp and W. Stout, "Almost sure invariance principles for partial sums of weakly dependent random variables," in *Memoirs AMS*, vol. 2, issue 2, no. 161, 1975.

[10] P. C. Shields, "String matching bounds via coding," *Ann. Prob.*, vol. 25, pp. 329–336, 1997.

[11] J. Shtarkov, "Coding of discrete sources with unknown statistics," in *Topics in Information Theory* (Coll. Math. Soc. J. Bolyai, no. 16), I. Csiszár and P. Elias, Eds. Amsterdam, The Netherlands: North Holland, 1977, pp. 559–574.

[12] A. A. Yushkevich, "On limit theorems connected with the concept of the entropy of Markov chains," *Usp. Mat. Nauk* vol. 8, pp. 177–180, 1953 (in Russian).