Optimal Lossless Compression: Source Varentropy and Dispersion

Ioannis Kontoyiannis
Department of Informatics
Athens U. of Economics & Business
Athens, Greece
Email: yiannis@aueb.gr

Sergio Verdú
Department of Electrical Engineering
Princeton University
Princeton, NJ, USA
Email: verdu@princeton.edu

***Abstract*—This work[1] deals with the fundamental limits of strictly-lossless variable-length compression of known sources without prefix constraints. The *source dispersion* characterizes the time-horizon over which it is necessary to code in order to approach the entropy rate within a pre-specified tolerance. We show that for a large class of sources, the dispersion of the source is equal to the *varentropy rate*, defined as the asymptotic per-symbol variance of the information random variables. We focus on ergodic Markov chains, whose optimal encodings are shown to be asymptotically normal and to satisfy an appropriate laws of the iterated logarithm.**

***Keywords* — Lossless data compression; source coding; fundamental limits; entropy rate; Markov sources; minimal source coding rate.**

I. Introduction

For a random process $\mathbf{X} = \{P_{X^n}\}_{n=1}^{\infty}$, assumed for simplicity to take values in a finite alphabet $\mathcal{A}$, the minimum asymptotically achievable source coding rate is the entropy rate,

$$H(\mathbf{X}) = \lim_{n\to\infty} \frac{1}{n} H(X^n) \tag{1}$$

$$= \lim_{n\to\infty} \frac{1}{n} \mathbb{E}[\imath_{X^n}(X^n)], \tag{2}$$

where $X^n = (X_1, X_2, \ldots, X_n)$ and the information of a random variable Z with distribution P_Z is defined as,

$$\imath_Z(a) = \log \frac{1}{P_Z(a)}, \tag{3}$$

under the following assumptions:

1) *Almost-lossless n-to-k fixed-length data compression:* Provided that the source is stationary and ergodic and the encoding failure probability does not exceed $0 < \epsilon < 1$, the minimum achievable rate $\frac{k}{n}$ is given by (1) as $n \to \infty$. This is a direct consequence of the Shannon-MacMillan theorem [10]. Dropping the assumption of stationarity/ergodicity, the fundamental limit is the limsup in probability of the normalized informations [2].
2) *Strictly lossless variable-length prefix data compression:* Provided that the limit in (1) exists (for which stationarity is sufficient) the minimal *average* source coding rate converges to (1). This is a consequence of the fact that for prefix codes the average encoded length cannot be smaller than the entropy [11], and the minimal average encoded length (achieved by the Huffman code), never exceeds the entropy plus one bit. If the limit in (1) does not exist, then the asymptotic minimal average source coding rate is simply the $\limsup$ of the normalized entropies [2]. For stationary ergodic sources, the source coding rate achieved by any prefix code is asymptotically almost surely bounded below by the entropy rate as a result of Barron's lemma [1], a bound which is achieved by the Shannon code.
3) *Strictly lossless variable-length data compression:* Stored files do not rely on prefix constraints to determine the boundaries between files. Instead, a pointer directory contains the starting and ending locations of the sequence of blocks occupied by each file in the storage medium (e.g. [15]). If no prefix constraints are imposed and the source is stationary and ergodic, the (random) rate of the optimum code converges in probability to (1). One way to reach this conclusion is the equivalence of the fundamental limits of almost-lossless fixed-length compression and strictly-lossless variable-length compression without prefix constraints [20].

It is of great practical interest to refine the above results to gauge the speed at which these limits are reached or, more ambitiously, the fundamental limit as a function of blocklength. We review some highlights of progress in that direction. To that end, we define the following key quantity:

Definition 1: The *varentropy rate* of a random process $\mathbf{X} = \{P_{X^n}\}_{n=1}^{\infty}$ is:

$$\sigma^2(\mathbf{X}) = \limsup_{n\to\infty} \frac{1}{n} \mathsf{Var}(\imath_{X^n}(X^n)). \tag{4}$$

[1] IK's work is co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program *"Education and Lifelong Learning* of the National Strategic Reference Framework (NSRF) Research Funding Program *THALES: Investing in knowledge society through the European Social Fund.* SV is supported in part by NSF Grant CCF-1016625 and by the *Center for Science of Information*, an NSF Science and Technology Center, under Grant CCF-0939370.

1) *Almost-lossless fixed-length source codes:* Refining a more general result of Yushkevich [22], Strassen [16] claimed (see the discussion in [8] regarding Strassen's proof) the following Gaussian approximation of the minimal rate compatible with error probability ϵ at blocklength n for non-equiprobable memoryless sources such that $\imath_X(X)$ is non-lattice:

$$\begin{aligned} R^*(n,\epsilon) &= H(\mathbf{X}) + \frac{\sigma(\mathbf{X})}{\sqrt{n}} Q^{-1}(\epsilon) \\ &\quad - \frac{1}{2n}\log_2\left(2\pi\sigma^2(\mathbf{X}) n e^{(Q^{-1}(\epsilon))^2}\right) \\ &\quad + \frac{\mu_3}{6\sigma^2(\mathbf{X})n}\left((Q^{-1}(\epsilon))^2 - 1\right) + o\left(\frac{1}{n}\right). \quad (5) \end{aligned}$$

Here and throughout, $R^*(n,\epsilon)$ denotes the lowest rate R such that the compression rate of the best code exceeds R with probability not greater than ϵ; the standard Gaussian tail function is $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-t^2/2}\,dt$; and μ_3 is the third centered absolute moment of $\imath_X(X)$. Apparently unaware of [16], Hayashi [3] established (5) but only up to order $o(n^{-1/2})$. An exact expression for the average error probability achieved by random binning is given in [21].

2) *Variable-length prefix codes:* The per-symbol redundancy (average length minus the entropy) of the Huffman codes is positive and behaves as $O\left(\frac{1}{n}\right)$. For biased coin flips, Szpankowski [17], [18] gave a precise study of the asymptotic behavior of the redundancy. Going beyond the analysis of the expected length, Kontoyiannis [6] gives a different kind of Gaussian approximation for the codelengths $\ell(\mathsf{f}_n(X^n))$ of arbitrary prefix codes f_n on memoryless data X^n, showing that, with probability one,[2] $\ell(\mathsf{f}_n(X^n))$ is eventually bounded below by a random variable that has an approximately Gaussian distribution,

$$\ell(\mathsf{f}_n(X^n)) \geq Z_n \text{ where } Z_n \overset{\mathcal{D}}{\approx} N(nH, n\sigma^2(\mathbf{X})), \quad (7)$$

Therefore, the codelengths $\ell(\mathsf{f}_n(X^n))$ will have at least Gaussian fluctuations of $O(\sqrt{n})$; this is further sharpened in [6] to a corresponding law of the iterated logarithm, stating that, with probability one, the compressed lengths $\ell(\mathsf{f}_n(X^n))$ will have fluctuations of $O(\sqrt{n \ln\ln n})$, infinitely often; with probability one:

$$\limsup_{n\to\infty} \frac{\ell(\mathsf{f}_n(X^n)) - H(X^n)}{\sqrt{2n\ln\ln n}} \geq \sigma(\mathbf{X}). \quad (8)$$

3) *Variable-length codes without prefix constraints:* Szpankowski and Verdú [19] show that for memoryless sources with marginal distribution P_X the best achievable average coding rate $\bar{R}(n)$ behaves as,

$$\bar{R}(n) = H(\mathbf{X}) - \frac{1}{2n}\log_2(8\pi e\sigma^2(\mathbf{X})n) + o\left(\frac{1}{n}\right), \quad (9)$$

if X is not equiprobable and $\imath_X(X)$ is a non-lattice random variable. Otherwise, they establish the redundancy is equal to $-\frac{1}{2n}\log_2 n + O(1)$.
An exact parametric expression of, as well as bounds on, the finite-blocklength fundamental limit are given by Verdú and Kontoyiannis [21]. They also show that, non-asymptotically, the prefix code designed to minimize the probability that the length exceeds a given bound incurs in a one bit penalty (Huffman and Shannon codes incur in higher penalties) relative to the optimal code not subject to the prefix constraint.

The remainder of the paper is organized as follows. Section III collects several observations on the varentropy rate as well as the central limit theorem and law of the iterated logarithm for the normalized informations for Markov chains. Section IV defines the source dispersion D as the limiting normalized variance of the optimal codelengths. In effect, the dispersion gauges the time one must wait for the source realization to become typical within a given tolerance. We identify a general class of sources, including Markov chains of any order, for which dispersion is equal to varentropy. In Section II we revisit the refined asymptotic results (7) and (8) of [6], and show that they remain valid for general (not necessarily prefix) compressors, and for a broad class of possibly infinite-memory sources. Section V examines in detail the finite-blocklength behavior of the fundamental limit $R^*(n,\epsilon)$ for the case of Markov sources. We prove tight, non-asymptotic and easily computable bounds for $R^*(n,\epsilon)$; these imply the following approximation:

Gaussian approximation: For every ergodic Markov source, the best achievable rate $R^*(n,\epsilon)$ satisfies,

$$nR^*(n,\epsilon) = nH(\mathbf{X}) + \sqrt{n}\sigma(\mathbf{X})Q^{-1}(\epsilon) + O(\log n). \quad (10)$$

Because of space limitations, results are stated without proof; further details and complete proofs can be found in [8].

II. Pointwise Asymptotics

In this section we examine the asymptotic behavior of the normalized difference between the codelength and the information (sometimes known as the pointwise redundancy).

Theorem 1: For any discrete source and any divergent deterministic sequence κ_n such that,

$$\lim_{n\to\infty} \frac{\log n}{\kappa_n} = 0, \quad (11)$$

we have:

(a) For any sequence $\{\mathsf{f}_n\}$ of codes, w. p. 1

$$\liminf_{n\to\infty} \frac{1}{\kappa_n}\left(\ell(\mathsf{f}_n(X^n)) - \imath_{X^n}(X^n)\right) \geq 0. \quad (12)$$

[2] A fixed-to-variable compressor for n-strings from finite alphabet $\mathcal{A}$ is simply an injective function,

$$\mathsf{f}_n \colon \mathcal{A}^n \to \{0,1\}^* = \{\varnothing, 0, 1, 00, 01, 10, 11, 000, \ldots\}. \quad (6)$$

A block of n symbols $a^n = (a_1, a_2, \ldots, a_n) \in \mathcal{A}^n$ is losslessly compressed by f_n into a binary string whose length is $\ell(\mathsf{f}_n(a^n))$ bits, where the length of any string $a \in \{0,1\}^*$ is denoted by $\ell(a)$.

(b) The sequence of optimal codes $\{\mathsf{f}_n^*\}$ achieves w.p. 1

$$\lim_{n\to\infty} \frac{1}{\kappa_n} \left(\ell(\mathsf{f}_n^*(X^n)) - \imath_{X^n}(X^n)\right) = 0. \tag{13}$$

As noted in the introduction for any discrete stationary ergodic process $\mathbf{X}$ the rate of the optimal code f_n^* converges in probability to the entropy rate. The next result shows that this holds almost surely. Moreover, no compressor can beat the entropy rate asymptotically with positive probability.

Theorem 2: Suppose that $\mathbf{X}$ is a stationary ergodic source with entropy rate $H(\mathbf{X})$.

(i) For any sequence $\{\mathsf{f}_n\}$ of codes,

$$\liminf_{n\to\infty} \frac{1}{n}\ell(\mathsf{f}_n(X^n)) \geq H(\mathbf{X}), \quad \text{w.p.1.} \tag{14}$$

(ii) The sequence of optimal codes $\{\mathsf{f}_n^*\}$ achieves,

$$\lim_{n\to\infty} \frac{1}{n}\ell(\mathsf{f}_n^*(X^n)) = H(\mathbf{X}), \quad \text{w.p.1.} \tag{15}$$

Theorem 2 is a simple consequence of Theorem 1 combined with the Shannon-Macmillan-Breiman theorem. The corresponding results for prefix codes were established in [1], [5], [6].

III. Varentropy

Some observations on the varentropy rate $\sigma^2(\mathbf{X})$, (also called the *minimal coding variance* in [6]) introduced in Definition 1:

1) If $\mathbf{X}$ is a stationary memoryless process with marginal distribution P_X, then the varentropy rate of $\mathbf{X}$ is equal to the varentropy of X, namely, the variance of the random variable $\imath_X(X)$, and it is zero if and only if X is equiprobable on its support.
2) In contrast to the first moment, we do not know whether stationarity is sufficient for $\limsup = \liminf$ in (4).
3) While the entropy-rate of a Markov chain admits a two-letter expression, the varentropy does not. In particular, if $\sigma^2(a)$ denotes the varentropy of $P_{X'|X}(\cdot\,|\,a)$, then the varentropy of the chain is, in general, not given by $\mathbb{E}[\sigma^2(X_0)]$.
4) The varentropy rate of Markov sources is typically nonzero. For example, for a first order Markov chain it was observed in [22], [7] that $\sigma^2(\mathbf{X}) = 0$ if and only if the source satisfies the following *deterministic equipartition property*: Every string x^{n+1} that starts and ends with the same symbol, has probability (given $X_1 = x_1$) q^n, for some constant q.

We also recall the following asymptotic properties of the information random variables $\imath_{X^n}(X^n)$; see [14] and the references therein.

Theorem 3: Let $\mathbf{X}$ be a stationary ergodic finite-state Markov chain.

(i) The varentropy rate $\sigma^2(\mathbf{X})$ is equal to the corresponding $\liminf$ of the normalized variances in (4), and it is finite.

(ii) The normalized information random variables are asymptotically normal, in the sense that, as $n \to \infty$,

$$\frac{\imath_{X^n}(X^n) - H(X^n)}{\sqrt{n}} \to N(0, \sigma^2(\mathbf{X})), \text{ in distribution.}$$

(iii) The normalized information random variables satisfy a corresponding law of the iterated logarithm:

$$\limsup_{n\to\infty} \frac{\imath_{X^n}(X^n) - H(X^n)}{\sqrt{2n\ln\ln n}} = \sigma(\mathbf{X}), \quad \text{w.p.1} \tag{16}$$

$$\liminf_{n\to\infty} \frac{\imath_{X^n}(X^n) - H(X^n)}{\sqrt{2n\ln\ln n}} = -\sigma(\mathbf{X}), \quad \text{w.p.1} \tag{17}$$

IV. Source Dispersion

We denote by $\mathsf{f}_n^* : \mathcal{A}^n \to \{0,1\}^*$ the optimal compressor that assigns the elements of $\mathcal{A}^n$ ordered in decreasing probabilities to the elements in $\{0,1\}^*$ ordered lexicographically. (It is immaterial how ties are broken.) This optimal code f_n^* is independent of the design target, in that, e.g., it is the same regardless of whether we want to minimize average length or the probability that the encoded length exceeds 1 KB or 1 MB. In fact, the code f_n^* possesses the following strong stochastic (competitive) optimality property over any other code f_n that can be losslessly decoded:

$$\mathbb{P}[\ell(\mathsf{f}_n(X^n)) \geq k] \geq \mathbb{P}[\ell(\mathsf{f}_n^*(X^n)) \geq k], \quad \text{for all } k \geq 0. \tag{18}$$

In addition to the minimal average compression rate at a given blocklength,

$$\bar{R}(n) = \frac{1}{n} \min_{\mathsf{f}_n} \mathbb{E}[\ell(\mathsf{f}_n(X^n))], \tag{19}$$

it is of interest to analyze the distribution of the optimum code lengths. To that end we let $R^*(n,\epsilon)$ be the lowest rate R such that the compression rate of the best code exceeds R bits/symbol with probability no greater than ϵ:

$$\min_{\mathsf{f}_n} \mathbb{P}[\ell(\mathsf{f}_n(X^n)) > nR] \leq \epsilon. \tag{20}$$

Note that we use the same notation as in the fundamental limit for almost-lossless fixed-length compression since they are equal.

Also of interest is $n^*(R,\epsilon)$: The smallest blocklength at which compression at rate R is possible with probability at least $1-\epsilon$; in other words, the minimum n required for (20) to hold.

Consider the following operational definition:

Definition 2: The *dispersion* D of a source $\{P_{X^n}\}_{n=1}^\infty$ is:

$$D = \limsup_{n\to\infty} \frac{1}{n}\mathsf{Var}(\ell(\mathsf{f}_n^*(X^n))). \tag{21}$$

As we show in Theorem 5, for a broad class of sources, the dispersion D is equal to the source varentropy rate $\sigma^2(\mathbf{X})$ defined in (4). Moreover, in view of the Gaussian approximation bounds for $R^*(n,\epsilon)$ in Section V – and more generally, as long as a similar two-term Gaussian approximation in terms of the entropy rate and varentropy rate can be established up to $o(1/\sqrt{n})$ accuracy – we can conclude the following: By the definition of $n^*(R,\epsilon)$, the source blocklength n required for the compression rate to exceed $(1+\eta)H(\mathbf{X})$ with probability no greater than $\epsilon > 0$ is approximated by,

$$n^*((1+\eta)H,\epsilon) \approx \frac{\sigma^2(\mathbf{X})}{H^2(\mathbf{X})}\left(\frac{Q^{-1}(\epsilon)}{\eta}\right)^2 \tag{22}$$

$$= \frac{D}{H^2(\mathbf{X})}\left(\frac{Q^{-1}(\epsilon)}{\eta}\right)^2, \tag{23}$$

i.e., by the product of a factor that depends only on the source (through $H(\mathbf{X})$ and $\sigma^2(\mathbf{X})$), and a factor that depends only on the design requirements ϵ and η.

Example 1: Coin flips with bias p have varentropy,

$$\sigma^2(\mathbf{X}) = p(1-p)\log_2^2 \frac{1-p}{p}, \tag{24}$$

so the key parameter in (22) which characterizes the time horizon required for the source to become "typical" is,

$$\frac{D}{H^2(\mathbf{X})} = p(1-p)\left(\frac{\log_2 \frac{1-p}{p}}{h(p)}\right)^2 \tag{25}$$

where $h(\cdot)$ denotes the binary entropy function in bits.

Example 2: For a memoryless source whose marginal is the geometric distribution,

$$P_X(k) = q(1-q)^k, \quad k \geq 0, \tag{26}$$

the ratio of varentropy to squared entropy is,

$$\frac{\sigma^2(\mathbf{X})}{H^2(\mathbf{X})} = \frac{D}{H^2(\mathbf{X})} = (1-q)\left(\frac{\log_2(1-q)}{h(q)}\right)^2, \tag{27}$$

Example 3: A binary symmetric Markov chain with transition probability α has

$$H(\mathbf{X}) = h(\alpha) \tag{28}$$

$$\sigma^2(\mathbf{X}) = \alpha(1-\alpha)\log_2^2 \frac{1-\alpha}{\alpha}, \tag{29}$$

so not only is the ultimate asymptotic achievable rate the same as a Bernoulli source with bias α but the time horizon to approach with the optimal code is identical. Note that a near-optimal encoding strategy is to send X_1 unencoded followed by the codeword that encodes the transitions using the optimal code for the Bernoulli source with bias α.

Figure 1 compares the normalized dispersion to the entropy for the Bernoulli, geometric and Poisson distributions. We see that, as the source becomes more compressible (lower entropy per letter), the horizon over which we need to compress in order to squeeze most of the redundancy out of the source gets longer.

Definition 3: A source $\mathbf{X}$ taking values on the finite alphabet $\mathcal{A}$ is a *linear information growth* source if the probability of every string is either zero or is asymptotically lower bounded by an exponential, that is, if there is a finite constant A and an integer $N_0 \geq 1$ such that, for all $n \geq N_0$, every nonzero-probability string $x^n \in \mathcal{A}^n$ satisfies $\imath_{X^n}(x^n) \leq An$. It is easy to check that memoryless sources as well as irreducible aperiodic Markov chains are linear information growth sources.

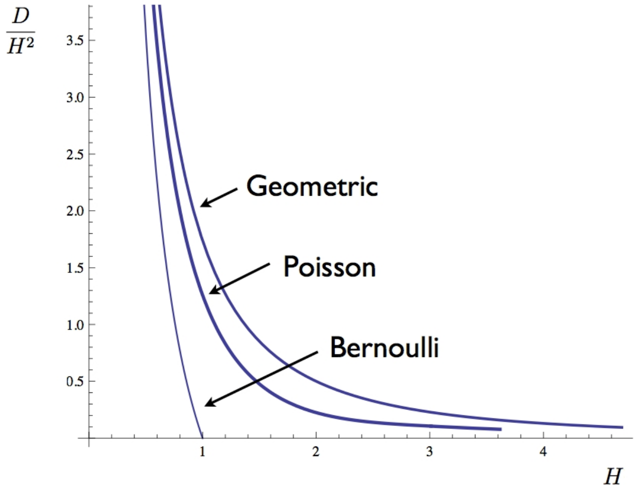

Fig. 1: Normalized dispersion as a function of entropy for memoryless sources

Linear information growth is sufficient for dispersion to equal varentropy:

Theorem 4: If the source has linear information growth, and finite varentropy, then, $D = \sigma^2(\mathbf{X})$.

V. Markov Sources

Theorem 5: Let $\mathbf{X}$ be an irreducible, aperiodic (not necessarily stationary) Markov source with entropy rate $H(\mathbf{X})$.

1) The varentropy rate $\sigma^2(\mathbf{X})$ defined in (4) exists as the limit,

$$\sigma^2(\mathbf{X}) = \lim_{n\to\infty}\frac{1}{n}\mathsf{Var}(\imath_{X^n}(X^n))). \tag{30}$$

2) The dispersion D defined in (21) exists as the limit,

$$D = \lim_{n\to\infty}\frac{1}{n}\mathsf{Var}(\ell(\mathsf{f}_n^*(X^n))). \tag{31}$$

3) $D = \sigma^2(\mathbf{X})$.
4) Suppose the varentropy rate (or, equivalently, the dispersion) is nonzero. Then it can be characterized in terms of the best achievable rate $R^*(n,\epsilon)$ as:

$$\sigma^2(\mathbf{X}) = \lim_{\epsilon\to 0}\lim_{n\to\infty}\frac{n\left(R^*(n,\epsilon) - H(\mathbf{X})\right)^2}{2\ln\frac{1}{\epsilon}}. \tag{32}$$

Next, we assume that the source is a stationary ergodic finite-alphabet (first-order) Markov chain, This enables us to analyze more precisely the behavior of the information random variables and, in particular, to invoke Theorem 3. Together with Theorem 1 particularized to $\kappa_n = \sqrt{n}$, we obtain

Theorem 6: Suppose $\mathbf{X}$ is a stationary ergodic Markov chain with entropy rate $H(\mathbf{X})$ and varentropy rate $\sigma^2(\mathbf{X})$. Then:

(i)

$$\frac{\ell(\mathsf{f}_n^*(X^n)) - H(\mathbf{X})}{\sqrt{n}} \longrightarrow N(0,\sigma^2(\mathbf{X})). \tag{33}$$

(ii) For any sequence of codes $\{\mathsf{f}_n\}$:

$$\limsup_{n\to\infty} \frac{\ell(\mathsf{f}_n(X^n)) - H(X^n)}{\sqrt{2n\ln\ln n}} \geq \sigma(\mathbf{X}), \text{ w.p.1}; \quad (34)$$

$$\liminf_{n\to\infty} \frac{\ell(\mathsf{f}_n(X^n)) - H(X^n)}{\sqrt{2n\ln\ln n}} \geq -\sigma(\mathbf{X}), \text{ w.p.1}. \quad (35)$$

(iii) The sequence of optimal codes $\{\mathsf{f}_n^*\}$ achieves the bounds in (34) and (35) with equality.

As far as the pointwise $\sqrt{n}$ and $\sqrt{2n\ln\ln n}$ asymptotics the optimal codelengths exhibit the same behavior as the Shannon prefix code and arithmetic coding. However, the large deviations behavior of the arithmetic and Shannon codes is considerably worse. Extensions to mixing sources with infinite memory are given in [8].

Theorem 7: Let $\epsilon \in (0, 1/2)$. Suppose $\mathbf{X}$ is an irreducible and aperiodic (not necessarily stationary) kth order Markov source with varentropy rate $\sigma^2(\mathbf{X}) > 0$. Then, there is a positive constant c_7 such that, for all n large enough,

$$nR^*(n,\epsilon) \leq nH(\mathbf{X}) + \sigma(\mathbf{X})\sqrt{n}Q^{-1}(\epsilon) + c_7, \quad (36)$$

Theorem 8: Under the same assumptions as in Theorem 7, for all n large enough,

$$nR^*(n,\epsilon) \geq nH(\mathbf{X}) + \sigma(\mathbf{X})\sqrt{n}Q^{-1}(\epsilon) - \frac{1}{2}\log_2 n - c_8 \quad (37)$$

where $c_8 > 0$ is a finite constant.

Remarks. **1.** Unlike the direct and converse coding theorems for memoryless sources in [21], the results of Theorems 7 and 8 are asymptotic in that we do not give explicit bounds for the constant terms. This is because the main probabilistic tool we use in [21] (the Berry-Esséen bound) does not have an equally precise counterpart for Markov chains. Specifically, in the proof of Theorem 9 we appeal to a Berry-Esséen bound established in [13], which does not give an explicit value for the multiplicative constant A.

2. If we restrict attention to the (much more narrow) class of *reversible* chains, then it is indeed possible to apply the Berry-Esséen bound of [9] to obtain explicit values for the constants in Theorems 7 and 8; but the resulting values are very loose, drastically limiting the engineering usefulness of the resulting bounds. Therefore, we have opted for the less explicit but much more general statements given above.

3. Similar comments to those in the last two remarks apply to the observation that Theorem 7 is a weaker bound than that established for for memoryless sources in [21], by a $(1/2)\log_2 n$ term. Instead of restricting our result to the much more narrow class of reversible chains, we chose to illustrate how this slightly weaker bound can be established in full generality, with a much simpler proof.

The proofs of Theorems 7 and 8 depend on a Berry-Esséen-type bound on the scaled information random variables

$$Z_n = \frac{\imath_{X^n}(X^n) - H(X^n)}{\sqrt{n}}, \quad (38)$$

Beyond the Shannon-McMillan-Breiman theorem, several refined asymptotic results have been established for this sequence; see, in particular, [16], [22], [4], [14] and the discussion in [7]. Unlike those results, we establish the following non-asymptotic bound.

Theorem 9: For an ergodic, kth order Markov source $\mathbf{X}$ with positive varentropy rate, there exists a finite constant $A > 0$ such that, for all $n \geq 1$,

$$\sup_{z\in\mathbb{R}} \left| \mathbb{P}\left[\imath_{X^n}(X^n) - nH(\mathbf{X}) > z\,\sigma(\mathbf{X})\sqrt{n}\right] - Q(z) \right| \leq \frac{A}{\sqrt{n}}.$$

References

[1] A. R. Barron, "*Logically smooth density estimation*," Ph. D. thesis, Dept. Electrical Engineering, Stanford University, Sep. 1985

[2] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Information Theory*, vol. 39, pp. 752–772, May 1993.

[3] M. Hayashi, "Second-order asymptotics in fixed-length source coding and intrinsic randomness," *IEEE Trans. on Information Theory,* vol. 54, no. 10, pp. 4619–4637, Oct. 2008

[4] I. A. Ibragimov, "Some limit theorems for stationary processes," *Theory Probability Applications*, vol. 7, pp. 349–382, 1962.

[5] J. C. Kieffer, "Sample converses in source coding theory," *IEEE Trans. on Information Theory*, vol. IT-37, no. 2, pp. 263–268, 1991.

[6] I. Kontoyiannis, "Second-order noiseless source coding theorems," *IEEE Trans. Information Theory*, vol. 43, no. 3, pp. 1339-1341, July 1997.

[7] I. Kontoyiannis. "Asymptotic recurrence and waiting times for stationary processes," *J. Theoretical Probability*, vol. 11, pp. 795-811, 1998.

[8] I. Kontoyiannis and S. Verdú, "Lossless data compression at finite blocklengths," arXiv 1212.2668, November 2012.

[9] B. Mann, *Berry-Esséen Central Limit Theorems for Markov Chains,* PhD thesis, Department of Mathematics, Harvard University, 1996.

[10] B. McMillan, "The basic theorems of information theory," *Annals of Mathematical Statististics*, vol. 24, no. 2, pp. 196–219, June 1953.

[11] B. McMillan, "Two inequalities implied by unique decipherability," *IRE Trans. Information Theory*, vol. 2, pp. 115–116, Dec. 1956.

[12] S. P. Meyn and R.L. Tweedie, *Markov Chains and Stochastic Stability*, Second edition, Cambridge University Press, 2009.

[13] S. V. Nagaev, "More exact limit theorems for homogeneous Markov chains," *Theory Probability Applications*, vol. 6, pp. 62-81, 1961.

[14] W. Philipp and W. Stout, *Almost Sure Invariance Principles for Partial Sums of Weakly Dependent Random Variables,* Memoirs of the AMS, 1975.

[15] G. Somasundaram and A. Shrivastava, eds. *Information Storage and Management: Storing, Managing, and Protecting Digital Information in Classic, Virtualized, and Cloud Environments*, John Wiley & Sons, 2012.

[16] V. Strassen, "Asymptotische Abschäzungen in Shannons Informationstheorie," *Trans. Third Prague Conf. Information Theory, on Statistics, Decision Functions, Random Processes* (Liblice, 1962), pages 689-723., Publ. House Czech. Acad. Sci., Prague, 1964.

[17] W. Szpankowski, "Asymptotic average redundancy of Huffman (and other) block codes, *IEEE Trans. Information Theory*, vol. 46, no. 7, pp. 2434–2443, Nov. 2000.

[18] W. Szpankowski, "Average redundancy for known sources: ubiquitous trees in source coding." In *Proceedings, Fifth Colloquium on Mathematics and Computer Science (Blaubeuren, 2008)*, Discrete Math. Theor. Comput. Sci. Proc. AI, pp. 19-58. 2008.

[19] W. Szpankowski and S. Verdú, "Minimum expected length of fixed-to-variable lossless compression without prefix constraints," *IEEE Trans. on Information Theory,* vol. 57, no. 7, pp. 4017–4025, July 2011.

[20] S. Verdú, "Teaching lossless data compression," *IEEE Information Theory Society Newsletter,* vol. 61, no. 1, pp. 18–19, April 2011

[21] S. Verdú and I. Kontoyiannis, "Lossless data compression rate: Asymptotics and non-asymptotics," *46th Annual Conference on Information Sciences and Systems*, Princeton University, Princeton, NJ, March 2012.

[22] A. A. Yushkevich, "On limit theorems connected with the concept of entropy of Markov chains", *Uspekhi Matematicheskikh Nauk*, 8:5(57), pp. 177-180, 1953