Optimal Lossless Data Compression: Non-Asymptotics and Asymptotics

Ioannis Kontoyiannis, *Fellow, IEEE*, and Sergio Verdú, *Fellow, IEEE*

***Abstract*— This paper provides an extensive study of the behavior of the best achievable rate (and other related fundamental limits) in variable-length strictly lossless compression. In the non-asymptotic regime, the fundamental limits of fixed-to-variable lossless compression with and without prefix constraints are shown to be tightly coupled. Several precise, quantitative bounds are derived, connecting the distribution of the optimal code lengths to the source information spectrum, and an exact analysis of the best achievable rate for arbitrary sources is given. Fine asymptotic results are proved for arbitrary (not necessarily prefix) compressors on general mixing sources. Nonasymptotic, explicit Gaussian approximation bounds are established for the best achievable rate on Markov sources. The source dispersion and the source varentropy rate are defined and characterized. Together with the entropy rate, the varentropy rate serves to tightly approximate the fundamental nonasymptotic limits of fixed-to-variable compression for all but very small block lengths.**

***Index Terms*— Lossless data compression, fixed-to-variable source coding, fixed-to-fixed source coding, entropy, finite-block length fundamental limits, central limit theorem, Markov sources, varentropy, minimal coding variance, source dispersion.**

I. Fundamental Limits

A. Asymptotics: Entropy Rate

For a random source $\mathbf{X} = \{P_{X^n}\}$, assumed for simplicity to take values in a finite alphabet $\mathcal{A}$, the minimum asymptotically achievable source coding rate (bits per source sample) 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* (in bits) of a random variable Z with distribution P_Z is defined as

$$\imath_Z(a) = \log_2 \frac{1}{P_Z(a)}, \tag{3}$$

The foregoing asymptotic fundamental limit holds under the following settings:

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 [25]. Dropping the assumption of stationarity/ergodicity, the fundamental limit is the lim-sup in probability of the normalized informations [13].
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 [26], 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 lim sup of the normalized entropies [13]. 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 [2], a bound which is achieved by the Shannon code.
3) *Strictly lossless variable-length data compression:* As we elaborate below, 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). As noted in [40], [41], prefix constraints at the level of the whole encoded file are superfluous in most applications. 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. [36]). It should be noted that notwithstanding his introduction of the Shannon code, in [34], Shannon never imposed prefix constraints, and considered coding of the whole file rather than symbol-by-symbol.

B. Nonasymptotics: Optimum Fixed-to-Variable Codes

A fixed-to-variable compressor for strings of length n from an alphabet $\mathcal{A}$ is an injective function $\mathsf{f}_n : \mathcal{A}^n \to \{0,1\}^*$ where

$$\{0,1\}^* = \{\varnothing, 0, 1, 00, 01, 10, 11, 000, 001, \ldots\}. \tag{4}$$

The length of a string $a \in \{0,1\}^*$ is denoted by $\ell(a)$. Therefore, a block (or string, or file) of n symbols

Manuscript received December 11, 2012; revised November 3, 2013; accepted November 5, 2013. Date of publication November 14, 2013; date of current version January 15, 2014. S. Verdú was supported in part by the National Science Foundation under Grant CCF- 1016625, in part by the Center for Science of Information, and in part by the NSF Science and Technology Center under Grant CCF-0939370. I. Kontoyiannis was supported by the European Union and Greek National Funds through the Operational Program Education and Lifelong Learning of the National Strategic Reference Framework Research Funding Program THALES: Investing in knowledge society through the European Social Fund.

I. Kontoyiannis is with the Department of Informatics, Athens U of Econ and Business, Athens 10675, Greece (e-mail: yiannis@aueb.gr).

S. Verdú is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: verdu@princeton.edu).

Communicated by T. Weissman, Associate Editor for Shannon Theory.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2013.2291007

0018-9448 © 2013 IEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

$a^n = (a_1, a_2, \ldots, a_n) \in \mathcal{A}^n$ is losslessly compressed by f_n into a binary string whose length is $\ell(\mathsf{f}_n(a^n))$ bits.

If the file $X^n = (X_1, X_2, \ldots, X_n)$ to be compressed is generated by a probability law P_{X^n}, a basic information-theoretic object of study is the distribution of the rate of the optimal compressor, as a function of the blocklength n and the distribution P_{X^n}. The best achievable compression performance at finite blocklengths is characterized by the following fundamental limits:

1) $R^*(n, \epsilon)$ is the lowest rate R such that the compression rate of the best code exceeds R bits/symbol with probability not greater than $0 \leq \epsilon < 1$:

$$\min_{\mathsf{f}_n} \mathbb{P}[\ell(\mathsf{f}_n(X^n)) > nR] \leq \epsilon. \tag{5}$$

2) $\epsilon^*(n, k)$ is the best achievable excess-rate probability, namely, the smallest possible probability that the compressed length is greater than or equal to k:

$$\epsilon^*(n, k) = \min_{\mathsf{f}_n} \mathbb{P}[\ell(\mathsf{f}_n(X^n)) \geq k]. \tag{6}$$

3) $n^*(R, \epsilon)$ is 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 (5) to hold.

4) $\bar{R}(n)$ is the minimal average compression rate:

$$\bar{R}(n) = \frac{1}{n} \min_{\mathsf{f}_n} \mathbb{E}[\ell(\mathsf{f}_n(X^n))]. \tag{7}$$

Naturally, the fundamental limits in 1), 2) and 3) are equivalent in the sense that knowledge of one of them (as a function of its parameters) determines the other two. For example,

$$R^*(n, \epsilon) = \frac{k}{n} \quad \text{iff } \epsilon^*(n, k+1) \leq \epsilon < \epsilon^*(n, k). \tag{8}$$

The minima in the fundamental limits (5), (6), (7) are achieved by an optimal compressor f_n^* that assigns the elements of $\mathcal{A}^n$ ordered in decreasing probabilities to the elements in $\{0, 1\}^*$ ordered lexicographically as in (4). In particular, $R^*(n, \epsilon)$ is the smallest R such that

$$\mathbb{P}[\ell(\mathsf{f}_n^*(X^n)) > nR] \leq \epsilon. \tag{9}$$

As for 4), we observe that (8) results in:

$$\begin{aligned} \bar{R}(n) &= \frac{1}{n} \mathbb{E}[\ell(\mathsf{f}_n^*(X^n))] && (10) \\ &= \frac{1}{n} \sum_{k=1}^{\infty} \epsilon^*(n, k) && (11) \\ &= \int_0^1 R^*(n, x)\, dx && (12) \end{aligned}$$

We emphasize that the optimum 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 compressor f_n:

$$\mathbb{P}[\ell(\mathsf{f}_n(X^n)) \geq k] \geq \mathbb{P}[\ell(\mathsf{f}_n^*(X^n)) \geq k], \;\; k \geq 0. \tag{13}$$

An optimal compressor f_n^* must satisfy the following constructive property.

Property 1: Let $k_n = \lfloor \log_2(1 + |\mathcal{A}|^n) \rfloor$. For $k = 1, 2, \ldots, k_n$, any optimal code f_n^* assigns strings of length $0, 1, 2, \ldots, k-1$ to each of the

$$1 + 2 + 4 + \cdots + 2^{k-1} = 2^k - 1 \tag{14}$$

most likely elements of $\mathcal{A}^n$. If $\log_2(1 + |\mathcal{A}|^n)$ is not an integer, then f_n^* assigns strings of length k_n to the $|\mathcal{A}|^n + 1 - 2^{k_n}$ least likely elements of $\mathcal{A}^n$.

According to Property 1, the length of the longest codeword assigned by f_n^* is $\lfloor n \log_2 |\mathcal{A}| \rfloor$. To check this, note that if $\log_2(1 + |\mathcal{A}|^n)$ is an integer, then the maximal length is $\log_2(1 + |\mathcal{A}|^n) - 1 = \lfloor n \log_2 |\mathcal{A}| \rfloor$, otherwise, the maximal length is $\lfloor \log_2(1 + |\mathcal{A}|^n) \rfloor = \lfloor n \log_2 |\mathcal{A}| \rfloor$.

Note that Property 1 is a necessary and sufficient condition for optimality but it does not determine f_n^* uniquely: Not only does it not specify how to break ties among probabilities but any swap between two codewords of the same length preserves optimality. As in the following example, it is convenient, however, to think of f_n^* as the unique compressor constructed by breaking ties lexicographically and by assigning the elements of $\{0, 1\}^*$ in the lexicographic order of (4). Then, if a^n is the kth most probable outcome its encoded version has length

$$\ell(\mathsf{f}_n^*(a^n)) = \lfloor \log_2 k \rfloor \tag{15}$$

Example 1: Suppose $n = 4$, $\mathcal{A} = \{\circ, \bullet\}$, and the source is memoryless with $\mathbb{P}[X = \bullet] > \mathbb{P}[X = \circ]$. Then the following compressor is optimal:

$$\begin{aligned}
\mathsf{f}_4^*(\bullet\ \bullet\ \bullet\ \bullet) &= \varnothing \\
\mathsf{f}_4^*(\bullet\ \bullet\ \bullet\ \circ) &= \mathtt{0} \\
\mathsf{f}_4^*(\bullet\ \bullet\ \circ\ \bullet) &= \mathtt{1} \\
\mathsf{f}_4^*(\bullet\ \circ\ \bullet\ \bullet) &= \mathtt{00} \\
\mathsf{f}_4^*(\circ\ \bullet\ \bullet\ \bullet) &= \mathtt{01} \\
\mathsf{f}_4^*(\circ\ \circ\ \bullet\ \bullet) &= \mathtt{10} \\
\mathsf{f}_4^*(\circ\ \bullet\ \bullet\ \circ) &= \mathtt{11} \\
\mathsf{f}_4^*(\circ\ \bullet\ \circ\ \bullet) &= \mathtt{000} \\
\mathsf{f}_4^*(\bullet\ \bullet\ \circ\ \circ) &= \mathtt{001} \\
\mathsf{f}_4^*(\bullet\ \circ\ \circ\ \bullet) &= \mathtt{010} \\
\mathsf{f}_4^*(\bullet\ \circ\ \bullet\ \circ) &= \mathtt{011} \\
\mathsf{f}_4^*(\circ\ \circ\ \circ\ \bullet) &= \mathtt{100} \\
\mathsf{f}_4^*(\circ\ \circ\ \bullet\ \circ) &= \mathtt{101} \\
\mathsf{f}_4^*(\circ\ \bullet\ \circ\ \circ) &= \mathtt{110} \\
\mathsf{f}_4^*(\bullet\ \circ\ \circ\ \circ) &= \mathtt{111} \\
\mathsf{f}_4^*(\circ\ \circ\ \circ\ \circ) &= \mathtt{0000}.
\end{aligned}$$

Although f_n^* is not a prefix code, the decompressor is able to recover the source file a^n exactly from $\mathsf{f}_n^*(a^n)$ and its knowledge of n and P_{X^n}. Since the whole source file is compressed, it is not necessary to impose a prefix condition in order for the decompressor to know where the compressed file starts and ends. One of the goals of this paper is to analyze how much compressed efficiency can be improved in strictly lossless source coding by dropping the the prefix-free constraint at the block level.

C. Nonasymptotics: Optimum Fixed-to-Variable Prefix Codes

In this subsection we deal with the case when the prefix condition is imposed at the level of the whole encoded file, which is more efficient than imposing it at the level of each symbol–a well-studied setup which does not exploit memory and which we do not deal with in this paper. The fixed-to-variable prefix code that minimizes the average length is the Huffman code, achieving the average compression rate $\bar{R}_{\mathsf{p}}(n)$ (which is strictly larger than $\bar{R}(n)$), defined as in (7) but restricting the minimization to prefix codes.

Also of interest is to investigate the optimum rate of the prefix code that minimizes the probability that the length exceeds a given threshold[1]; if the minimization in (5) is carried out with respect to codes that satisfy the prefix condition, then the corresponding fundamental limit is denoted by $R_{\mathsf{p}}(n, \epsilon)$, and analogously $\epsilon_{\mathsf{p}}(n, k)$ for (6). Note that the optimum prefix code achieving those fundamental limits will, in general, depend on k. The following new result shows that the corresponding fundamental limits, with and without the prefix condition, are tightly coupled:

Theorem 1: Suppose all elements in $\mathcal{A}$ have positive probability. For all $n \geq 1$:

1) For each integer $k \geq 1$:

$$\epsilon_{\mathsf{p}}(n, k+1) = \begin{cases} \epsilon^*(n, k) & k < n \log_2 |\mathcal{A}| \\ 0 & k \geq n \log_2 |\mathcal{A}|. \end{cases} \quad (16)$$

2) If $|\mathcal{A}|$ is not a power of 2, then for $0 \leq \epsilon < 1$:

$$R_{\mathsf{p}}(n, \epsilon) = R^*(n, \epsilon) + \frac{1}{n}. \quad (17)$$

If $|\mathcal{A}|$ is a power of 2, then (17) holds for $\epsilon \geq \min_{a^n \in \mathcal{A}^n} P_{X^n}(a^n)$, while we have,

$$R_{\mathsf{p}}(n, \epsilon) = R^*(n, \epsilon) = \log_2 |\mathcal{A}| + \frac{1}{n}, \quad (18)$$

for $0 \leq \epsilon < \min_{a^n \in \mathcal{A}^n} P_{X^n}(a^n)$.

Proof: 1): Fix k and n satisfying $2^k < |\mathcal{A}|^n$. Since there is no benefit in assigning shorter lengths, any Kraft-inequality-compliant code $\mathsf{f}_n^{\mathsf{p}}$ that minimizes $\mathbb{P}[\ell(\mathsf{f}_n(X^n)) > k]$ assigns strings of length k to each of the $2^k - 1$ largest masses of P_{X^n}. Assigning to all the other elements of $\mathcal{A}^n$ binary strings of length equal to

$$\ell_{\max} = \lceil k + \log_2(|\mathcal{A}|^n - 2^k + 1) \rceil \quad (19)$$

guarantees that the Kraft sum is satisfied. On the other hand, according to Property 1, the optimum code f_n^* without prefix constraints encodes each of the $2^k - 1$ largest masses of P_{X^n} with strings of lengths ranging from 0 to $k - 1$. Therefore,

$$\mathbb{P}[\ell(\mathsf{f}_n^{\mathsf{p}}(X^n)) \geq k+1] = \mathbb{P}[\ell(\mathsf{f}_n^*(X^n)) \geq k], \quad (20)$$

or, equivalently, $\epsilon_{\mathsf{p}}(n, k+1) = \epsilon^*(n, k)$.

If $2^k \geq |\mathcal{A}|^n$, then a zero-error n-to-k code exists, and hence $\epsilon_{\mathsf{p}}(n, k+1) = 0$.

2): For brevity, let $p_{\min} = \min_{a^n \in \mathcal{A}^n} P_{X^n}(a^n)$. From Property 1 it follows that if $|\mathcal{A}|$ is not a power of 2, then

$$\epsilon^*(n, \lceil n \log_2 |\mathcal{A}| \rceil) = 0, \quad (21)$$

while if $|\mathcal{A}|$ is a power of 2, then

$$\epsilon^*(n, \lceil n \log_2 |\mathcal{A}| \rceil + 1) = 0 \quad (22)$$

$$\epsilon^*(n, \lceil n \log_2 |\mathcal{A}| \rceil) = p_{\min} \quad (23)$$

On the other hand, 1) implies that:

$$\epsilon_{\mathsf{p}}(n, \lceil n \log_2 |\mathcal{A}| \rceil + 1) = 0 \quad (24)$$

$$\begin{aligned} \epsilon_{\mathsf{p}}(n, \lceil n \log_2 |\mathcal{A}| \rceil) &= \epsilon^*(n, \lceil n \log_2 |\mathcal{A}| \rceil - 1) \\ &\geq p_{\min}. \end{aligned} \quad (25)$$

Furthermore, $R_{\mathsf{p}}(n, \cdot)$ can be obtained from $\epsilon_{\mathsf{p}}(n, \cdot)$ analogously to (8):

$$R_{\mathsf{p}}(n, \epsilon) = \frac{i}{n} \text{ iff } \epsilon_{\mathsf{p}}(n, i) \leq \epsilon < \epsilon_{\mathsf{p}}(n, i-1). \quad (26)$$

Together with (8) and (16), (26) implies that (17) holds if $\epsilon \geq p_{\min}$.

If $\epsilon < p_{\min}$, then (21)–(25) result in (18) when $|\mathcal{A}|$ is a power of 2. If $|\mathcal{A}|$ is not a power of 2 and $0 \leq \epsilon < p_{\min}$, then,

$$R^*(n, \epsilon) = \frac{\lceil n \log_2 |\mathcal{A}| \rceil}{n} \quad (27)$$

$$R_{\mathsf{p}}(n, \epsilon) = \frac{\lceil n \log_2 |\mathcal{A}| \rceil + 1}{n}, \quad (28)$$

completing the proof. ■

D. Nonasymptotics: Optimum Fixed-to-Fixed Almost-Lossless Codes

As pointed out in [40], [41], the quantity $\epsilon^*(n, k)$ is intimately related to the problem of almost-lossless fixed-to-fixed data compression. Assume the nontrivial compression regime in which $2^k < |\mathcal{A}|^n$. The optimal n-to-k fixed-to-fixed compressor assigns a unique string of length k to each of the $2^k - 1$ most likely elements of $\mathcal{A}^n$, and assigns all the others to the remaining binary string of length k, which signals an *encoder failure.* Thus, the source strings that are decodable error-free by the optimal n-to-k scheme are precisely those that are encoded with lengths ranging from 0 to $k - 1$ by the optimum variable-length code (Property 1). Therefore, $\epsilon^*(n, k)$, defined in (6) as a fundamental limit of (strictly) lossless variable-length codes is, in fact, equal to the minimum error probability of an n-to-k code. Accordingly, the results obtained in this paper apply to the standard paradigm of almost-lossless fixed-to-fixed compression as well as to the setup of lossless fixed-to-variable compression without prefix-free constraints at the block level.

The case $2^k \geq |\mathcal{A}|^n$ is rather trivial: The minimal probability of encoding failure for an n-to-k code is 0, which again coincides with $\epsilon^*(n, k)$, unless $|\mathcal{A}|^n = 2^k$, in which case, as we saw in (23),

$$\epsilon^*(n, k) = \min_{a^n \in \mathcal{A}^n} P_{X^n}(a^n). \quad (29)$$

[1] This was considered in [15] using the minimization of the moment generating function at a given argument as a proxy (see also [14].)

E. Existing Asymptotic Results

1) Optimal Variable-Length Codes: Based on the correspondence we just saw between optimal almost-lossless fixed-to-fixed codes and optimal strictly lossless fixed-to-variable codes, previous results on the asymptotics of fixed-to-fixed compression can be brought to bear. In particular the Shannon-MacMillan theorem [25], [34] implies that for a stationary ergodic finite-alphabet source with entropy rate $H(\mathbf{X})$, and for all $0 < \epsilon < 1$,

$$\lim_{n\to\infty} R^*(n,\epsilon) = H(\mathbf{X}). \tag{30}$$

Suppose X^n is generated by a memoryless source with distribution,

$$P_{X^n} = P_X \times P_X \times \cdots \times P_X, \tag{31}$$

where P_X has entropy $H(X)$ and varentropy[2]

$$\sigma^2(X) = \mathsf{Var}(\imath_X(X)). \tag{32}$$

For the expected length, Szpankowski and Verdú [38] show that the behavior of (7) for non-equiprobable sources is

$$\bar{R}(n) = H(X) - \frac{1}{2n}\log_2 n + O\left(\frac{1}{n}\right), \tag{33}$$

which is also refined to show that if $\imath_X(X)$ is non-lattice,[3] then,

$$\bar{R}(n) = H(X) - \frac{1}{2n}\log_2(8\pi e\sigma^2(X)n) + o\left(\frac{1}{n}\right). \tag{34}$$

Yushkevich [44] showed that

$$R^*(n,\epsilon) = H(X) + \frac{\sigma(X)}{\sqrt{n}}Q^{-1}(\epsilon) + o\left(\frac{1}{\sqrt{n}}\right) \tag{35}$$

a result that was refined for non-equiprobable memoryless sources such that $\imath_X(X)$ is non-lattice by Strassen [37]:[4]

$$\begin{aligned} R^*(n,\epsilon) = H(X) &+ \frac{\sigma(X)}{\sqrt{n}}Q^{-1}(\epsilon) \\ &- \frac{1}{2n}\log_2\left(2\pi\sigma^2(X)ne^{(Q^{-1}(\epsilon))^2}\right) \\ &+ \frac{\mu_3}{6\sigma^2(X)n}\left((Q^{-1}(\epsilon))^2 - 1\right) + o\left(\frac{1}{n}\right) \end{aligned} \tag{36}$$

where $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-t^2/2}\,dt$ denotes the standard Gaussian tail function, and μ_3 is the third centered absolute moment of $\imath_X(X)$.

Also in the context of memoryless sources with finite-alphabet $\mathcal{A}$ all whose letters have positive probabilitites, the exponential decrease of the error probability is given by (e.g. [9]) the parametric expression:

$$\lim_{n\to\infty}\frac{1}{n}\log\frac{1}{\epsilon^*(n,nR)} = D(P_{X_\alpha}||X) \tag{37}$$

$$R = H(X_\alpha) \tag{38}$$

where $R \in (H(X), \log|\mathcal{A}|)$, $\alpha \in (0,1)$, and

$$P_{X_\alpha}(a) = \frac{1}{c}\,P_X^\alpha(a) \tag{39}$$

with $c = \sum_{a\in\mathcal{A}} P_X^\alpha(a)$.

2) Optimal Prefix Variable-Length Codes: In contrast to (33), when a prefix-free condition is imposed, we have the well-known behavior for the average rate,

$$\bar{R}_{\mathsf{p}}(n) = H(\mathbf{X}) + O\left(\frac{1}{n}\right), \tag{40}$$

for any source for which the limit in (1) exists, see, e.g., [8].

It follows immediately from Theorem 1 that the prefix-free condition incurs no loss as far as the limit in (30) is concerned:

$$\lim_{n\to\infty} R_{\mathsf{p}}(n,\epsilon) = H(\mathbf{X}). \tag{41}$$

Kontoyiannis [19] 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, $\ell(\mathsf{f}_n(X^n))$ is eventually bounded below by a random variable that has an approximately Gaussian distribution,

$$\ell(\mathsf{f}_n(X^n)) \geq Z_n \;\text{ where }\; Z_n \overset{\mathcal{D}}{\approx} N(nH(X), n\sigma^2(X)); \tag{42}$$

and $\sigma^2(X)$ is the varentropy as in (32). Therefore, the codelengths $\ell(\mathsf{f}_n(X^n))$ will have at least Gaussian fluctuations of $O(\sqrt{n})$; this is further sharpened in [19] 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:

$$\limsup_{n\to\infty}\frac{\ell(\mathsf{f}_n(X^n)) - H(X^n)}{\sqrt{2n\ln\ln n}} \geq \sigma(X). \tag{43}$$

Both results (42) and (43) are shown to hold for Markov sources as well as for a wide class of mixing sources with infinite memory.

As far as the large deviations of the distribution of the lengths, [27] shows for a class of sources with memory that neither the prefix constraint nor universality of the compressor degrade the optimum error exponent, which is in fact achieved by the Lempel-Ziv compressor.

F. Outline of Main New Results

Theorem 1 shows that the prefix constraint only incurs one bit penalty as far as the distribution of the optimum rate is concerned. Note that requires the prefix code to be optimized for a given overflow probability. In contrast, as we commented in Sections I-E.1 and I-E.2, the penalty in average rate incurred by a hypothetical Huffman code that operated at the whole file level is larger.

Section II gives a general analysis of the distribution of the lengths of the optimal lossless code for any discrete information source. We adopt a single-shot approach which can be particularized to the conventional case in which the source produces a string of symbols of either deterministic or random length. In Theorems 3 and 4 we give simple achievability and converse bounds, showing that the distribution function of the optimal codelengths, $\mathbb{P}\left[\ell(\mathsf{f}^*(X)) \leq t\right]$, is intimately related to the distribution function of the information random variable,

[2] $\sigma^2(X)$ is called *minimal coding variance* in [19].

[3] A discrete random variable is *lattice* if all its masses are on a subset of some lattice $\{\nu + k\varsigma\ ;\ k \in \mathbb{Z}\}$.

[4] See the discussion in Section V regarding Strassen's claimed proof of this result.

$\mathbb{P}[\imath_X(X) \leq t]$. Also we observe that the optimal codelengths $\ell(\mathsf{f}^*(X))$ are always bounded above by $\imath_X(X)$, and we give an example where their distributions are noticeably different. However, significant deviations cannot be too probable according to Theorem 5.

Theorem 7 offers an exact, non-asymptotic expression for the best achievable rate $R^*(n, \epsilon)$. An exact expression for the average probability of error achieved by (almost-lossless) random binning, is given in Theorem 8.

General non-asymptotic and asymptotic results for the expected optimal length, $\bar{R}(n)$, are obtained in Section III.

In Section IV we revisit the refined asymptotic results (42) and (43) of [19], 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 memoryless sources. We prove tight, non-asymptotic and easily computable bounds for $R^*(n, \epsilon)$. combining the results of Theorems 17 and 18 implies the following approximation:

> *Gaussian approximation I:* For a memoryless source with entropy $H(X)$ and varentropy $\sigma^2(X)$, the best achievable rate $R^*(n, \epsilon)$ satisfies
>
> $$nR^*(n,\epsilon) \approx nH(X) + \sigma(X)\sqrt{n}Q^{-1}(\epsilon) - \frac{1}{2}\log_2 n + O(1) \tag{44}$$
>
> In view of Theorem 1, the same holds for $R_{\mathsf{p}}(n, \epsilon)$ in the case of prefix codes.

The approximation (44) is established by combining the general results of Section II with the classical Berry-Esséen bound [22], [30]. This approximation is made precise in a non-asymptotic way, and all the constants involved are explicitly identified.

In Section VI, achievability and converse bounds (Theorems 19 and 20) are established for $R^*(n, \epsilon)$, in the case of general ergodic Markov sources. Those results are analogous (though slightly weaker) to those in Section V. We also define the varentropy rate of an arbitrary source as the limiting normalized variance of the information random variables $\imath_{X^n}(X^n)$, and we show that, for Markov chains, it plays the same role as the varentropy defined in (32) for memoryless sources. Those results in particular imply the following:

> *Gaussian approximation II:* For any ergodic Markov source with entropy rate $H(\mathbf{X})$ and varentropy rate $\sigma^2(\mathbf{X})$, the blocklength $n^*(R, \epsilon)$ required for the compression rate to exceed $(1 + \eta)H(\mathbf{X})$ with probability no greater than $\epsilon > 0$, satisfies,
>
> $$n^*((1+\eta)H(\mathbf{X}), \epsilon) \approx \frac{\sigma^2(\mathbf{X})}{H^2(\mathbf{X})}\left(\frac{Q^{-1}(\epsilon)}{\eta}\right)^2. \tag{45}$$

Finally, Section VII 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 probability, as in (45) above, with D in place of $\sigma^2(\mathbf{X})$. For a large class of sources (including ergodic Markov chains of any order), the dispersion D is shown to equal the varentropy rate $\sigma^2(\mathbf{X})$ of the source.

II. Non-Asymptotics for Arbitrary Sources

In this section we analyze the best achievable compression performance on a completely general discrete random source. In particular, (except where noted) we do not necessarily assume that the alphabet is finite and we do not exploit the fact that in the original problem we are interested in compressing a block of n symbols. In this way we even encompass the case where the source string length is a priori unknown at the decompressor. Thus, we consider a given probability mass function P_X defined on an arbitrary finite alphabet $\mathcal{X}$, which may (but is not assumed to) consist of variable-length strings drawn from some alphabet. The results can then be particularized to the setting in Section I, letting $\mathcal{X} \leftarrow \mathcal{A}^n$ and $P_X \leftarrow P_{X^n}$. Conversely, we can simply let $n = 1$ in Section I to yield the setting in this section.

The best achievable rate $R^*(n, \epsilon)$ at blocklength $n = 1$ is abbreviated as $R^*(\epsilon) = R^*(1, \epsilon)$. By definition, $R^*(\epsilon)$ is the lowest R such that

$$\mathbb{P}[\ell(\mathsf{f}^*(X)) > R] \leq \epsilon, \tag{46}$$

which is equal to the quantile function[5] of the integer-valued random variable $\ell(\mathsf{f}^*(X))$ evaluated at $1 - \epsilon$.

A. Achievability Bound

Our goal is to express the distribution of the optimal codelengths $\ell(\mathsf{f}^*(X))$ in terms of the distribution of $\imath_X(X)$. The first result is the following simple and powerful relationship:

Theorem 2: Assume that the elements of $\mathcal{X}$ are integer-valued with decreasing probabilities: $P_X(1) \geq P_X(2) \geq \cdots$, and that f^* is the optimal code that assigns sequentially and lexicographically the shortest available string. Then, for all $a \in \mathcal{X}$[6]:

$$\ell(\mathsf{f}^*(a)) \leq \imath_X(a) \tag{47}$$

Proof: Fix $i = 1, 2, \ldots$. Since there are $i - 1$ outcomes at least as likely as i we must have[7]

$$P_X(i) \leq \frac{1}{i}. \tag{48}$$

as otherwise the sum of the probabilities would exceed 1. Taking $\log_2$ of (48), results in

$$\begin{aligned} \imath_X(i) &\geq \log_2 i & (49)\\ &\geq \lfloor \log_2 i \rfloor & (50)\\ &= \ell(\mathsf{f}^*(i)) & (51) \end{aligned}$$

where (51) follows as in (15). ■

[5] The quantile function $\mathcal{Q}$: $[0, 1] \to \mathbb{R}$ is the "inverse" of the cumulative distribution function F. Specifically, $\mathcal{Q}(\alpha) = \min\{x : F(x) = \alpha\}$ if the set is nonempty; otherwise α lies within a jump $\lim_{x \uparrow x_\alpha} F(x) < \alpha < F(x_\alpha)$ and we define $\mathcal{Q}(\alpha) = x_\alpha$.

[6] A legacy of the Kraft inequality mindset, the term "ideal codelength" is sometimes used for $\imath_X(X)$, which is neither a codelength nor ideal in view of (47). As we illustrate in Figure 1, the difference between both sides of (48) may be substantial.

[7] This inequality was used by Wyner [43] to show that for a positive-integer-valued random variable Z with decreasing probabilities $\mathbb{E}[\log Z] \leq H(Z)$.

Theorem 2 results in the following achievability result:

Theorem 3: [40] For any $a \ge 0$,

$$\mathbb{P}\left[\ell(\mathsf{f}^*(X)) \ge a\right] \le \mathbb{P}\left[\imath_X(X) \ge a\right]. \tag{52}$$

Proof: Since neither the actual choice of f^* nor the labeling of the elements of $\mathcal{X}$ can affect either side of (52), we are free to choose those specified in Theorem 2. Then, (52) follows immediately. ■

We note that neither Theorem 2 nor Theorem 3 require X to take values on a finite alphabet. Theorem 3 is the starting point for the achievability result for $R^*(n, \epsilon)$ established for Markov sources in Theorem 19.

B. Converse Bounds

In Theorem 4 we give a corresponding converse result; cf. [40]. It will be used later to obtain sharp converse bounds for $R^*(n, \epsilon)$ for memoryless and Markov sources, in Theorems 18 and 20, respectively.

Theorem 4: For any nonnegative integer k,

$$\sup_{\tau > 0} \left\{\mathbb{P}\left[\imath_X(X) \ge k + \tau\right] - 2^{-\tau}\right\} \le \mathbb{P}\left[\ell(\mathsf{f}^*(X)) \ge k\right]. \tag{53}$$

Proof: As in the proof of Theorem 3, we label the values taken by X as the positive integers in decreasing probabilities. Fix an arbitrary $\tau > 0$. Define:

$$\mathcal{L} = \{i \in \mathcal{X} \colon P_X(i) \le 2^{-k-\tau}\} \tag{54}$$

$$\mathcal{C} = \{1, 2, \ldots 2^k - 1\}. \tag{55}$$

Then, abbreviating $P_X(\mathcal{B}) = \sum_{i \in \mathcal{B}} P_X(i)$, $\mathcal{B} \subset \mathcal{X}$,

$$\begin{aligned}
\mathbb{P}\big[\imath_X(X) &\ge k + \tau\big] \\
&= P_X(\mathcal{L}) && (56)\\
&= P_X(\mathcal{L} \cap \mathcal{C}) + P_X(\mathcal{L} \cap \mathcal{C}^c) && (57)\\
&\le P_X(\mathcal{L} \cap \mathcal{C}) + P_X(\mathcal{C}^c) && (58)\\
&\le (2^k - 1)2^{-k-\tau} + P_X(\mathcal{C}^c) && (59)\\
&< 2^{-\tau} + \mathbb{P}\left[\lfloor \log_2 X \rfloor \ge k\right] && (60)\\
&= 2^{-\tau} + \mathbb{P}\left[\ell(\mathsf{f}^*(X)) \ge k\right], && (61)
\end{aligned}$$

where (61) follows in view of (51). ■

Next we give another general converse bound, similar to that of Theorem 4, where this time we directly compare the codelengths $\ell(\mathsf{f}(X))$ of an arbitrary compressor with the values of the information random variable $\imath_X(X)$. Whereas from (47) we know that $\ell(\mathsf{f}(X))$ is always smaller than $\imath_X(X)$, Theorem 5 says that it cannot be much smaller with high probability. This is a natural analog of the corresponding converse established for prefix compressors in [2], and stated as Theorem 6 below.

Applied to a finite-alphabet source, Theorem 5 is the key bound in the derivation of all the pointwise asymptotic results of Section IV, namely, Theorems 12, 13 and 14. It is also the main technical ingredient of the proof of Theorem 23 in Section VII stating that the source dispersion is equal to its varentropy.

Theorem 5: For any compressor f and any $\tau > 0$:

$$\mathbb{P}\left[\ell(\mathsf{f}(X)) \le \imath_X(X) - \tau\right] \le 2^{-\tau}\left(\lfloor \log_2 |\mathcal{X}| \rfloor + 1\right). \tag{62}$$

Proof: Letting $1\{A\}$ denote the indicator function of the event A, the probability in (62) can be bounded above as

$$\begin{aligned}
\mathbb{P}[\ell(\mathsf{f}(X)) &\le \imath_X(X) - \tau] \\
&= \sum_{x \in \mathcal{X}} P_X(x)\, 1\left\{P_X(x) \le 2^{-\tau-\ell(\mathsf{f}(x))}\right\} && (63)\\
&\le 2^{-\tau} \sum_{x \in \mathcal{X}} 2^{-\ell(\mathsf{f}(x))} && (64)\\
&\le 2^{-\tau} \sum_{j=0}^{\lfloor \log_2 |\mathcal{X}| \rfloor} 2^j 2^{-j}, && (65)
\end{aligned}$$

where the sum in (64) is maximized if f assigns a string of length $j + 1$ only if it also assigns all strings of length j. Therefore, (65) holds because that code contains all the strings of lengths $0, 1, \ldots, \lfloor \log_2 |\mathcal{X}| \rfloor - 1$ plus

$$|\mathcal{X}| - 2^{\lfloor \log_2 |\mathcal{X}| \rfloor} + 1 \le 2^{\lfloor \log_2 |\mathcal{X}| \rfloor} \tag{66}$$

strings of length $\lfloor \log_2 |\mathcal{X}| \rfloor$. ■

We saw in Theorem 1 that the optimum prefix code under the criterion of minimum excess length probability incurs a penalty of at most one bit. The following elementary converse is derived in [2], [19]; its short proof is included for completeness.

Theorem 6: For any prefix code f, and any $\tau > 0$:

$$\mathbb{P}[\ell(\mathsf{f}(X)) < \imath_X(X) - \tau] \le 2^{-\tau}. \tag{67}$$

Proof: We have, as in the proof of Theorem 5 leading to (64),

$$\begin{aligned}
\mathbb{P}[\ell(\mathsf{f}(X)) \le \imath_X(X) - \tau] &\le 2^{-\tau} \sum_{x \in \mathcal{X}} 2^{-\ell(\mathsf{f}(x))} && (68)\\
&\le 2^{-\tau}, && (69)
\end{aligned}$$

where (69) is Kraft's inequality. ■

C. Exact Fundamental Limit

The following result expresses the non-asymptotic data compression fundamental limit $R^*(\epsilon) = R^*(1, \epsilon)$ as a parametric function of the source information spectrum.

Theorem 7: For all $a \ge 0$, the exact minimum rate compatible with given excess-length probability satisfies,

$$R^*(\epsilon) = \left\lceil \log_2\left(1 + M(2^a)\right)\right\rceil - 1, \tag{70}$$

with

$$\epsilon = \mathbb{P}[\imath_X(X) \ge a], \tag{71}$$

where $M(\beta)$ denotes the number of masses with probability strictly larger than $\frac{1}{\beta}$, and which can be expressed as:

$$\begin{aligned}
M(\beta) &= \beta\, \mathbb{P}\left[\imath_X(X) < \log_2 \beta\right] \\
&\quad - \int_1^\beta \mathbb{P}\left[\imath_X(X) \le \log_2 t\right] dt. && (72)
\end{aligned}$$

Proof: As above, the values taken by X are labeled as the positive integers in order of decreasing probability. By the definition of $M(\cdot)$, for any positive integer i, and $a > 0$,

$$P_X(i) \le 2^{-a} \iff 1 + M(2^a) \le i, \tag{73}$$

Therefore,

$$\mathbb{P}\left[\log_2 X \geq \log_2(1 + M_X(2^a))\right] = \mathbb{P}\left[\imath_X(X) \geq a\right] \tag{74}$$

Moreover, because of (51), for all $b \geq 0$,

$$\begin{aligned} \mathbb{P}\left[\log_2 X \geq \lceil b \rceil\right] &= \mathbb{P}\left[\ell(\mathsf{f}^*_X(X)) \geq b\right] && (75)\\ &\leq \mathbb{P}\left[\log_2 X \geq b\right] && (76) \end{aligned}$$

Particularizing (75) to $b = \lceil \log_2(1 + M_X(2^a)) \rceil$ and letting $\epsilon < 1$ be given by (71) we see that

$$\mathbb{P}[\ell(\mathsf{f}^*(X)) > R] = \epsilon \tag{77}$$

if $R = \lceil \log_2(1 + M_X(2^a)) \rceil - 1$, and

$$\mathbb{P}[\ell(\mathsf{f}^*(X)) > R] > \epsilon \tag{78}$$

if $R < \lceil \log_2(1 + M_X(2^a)) \rceil - 1$. Therefore, (70) holds.

The proof of (72) follows a sequence of elementary steps:

$$\begin{aligned} M(\beta) &= \sum_{x \in \mathcal{X}} 1\left\{P_X(x) > \frac{1}{\beta}\right\} && (79)\\ &= \mathbb{E}\left[\frac{1\{P_X(X) > \frac{1}{\beta}\}}{P_X(X)}\right] && (80)\\ &= \int_0^\infty \mathbb{P}\left[\frac{1\{P_X(X) > \frac{1}{\beta}\}}{P_X(X)} > t\right] dt && (81)\\ &= \int_0^\beta \mathbb{P}\left[\frac{1}{\beta} < P_X(X) < \frac{1}{t}\right] dt && (82)\\ &= \int_0^\beta \mathbb{P}\left[\frac{1}{\beta} < P_X(X)\right] - \mathbb{P}\left[P_X(X) \geq \frac{1}{t}\right] dt && (83)\\ &= \beta\, \mathbb{P}\left[\imath_X(X) < \log_2 \beta\right] \\ &\quad - \int_1^\beta \mathbb{P}\left[\imath_X(X) \leq \log_2 t\right] dt. && (84) \end{aligned}$$

■

While Theorem 7 gives $R^*(\epsilon) = R^*(1, \epsilon)$ exactly for those ϵ which correspond to values taken by the complementary cumulative distribution function of the information random variable $\imath_X(X)$, a continuous sweep of $a > 0$ gives a very dense grid of values, unless X (whose alphabet size typically grows exponentially with n in the fixed-to-variable setup) takes values in a very small alphabet. From the value of a we can obtain the probability in the right side of (71). The optimum code achieves an excess probability $\epsilon = \mathbb{P}\left[\ell(\mathsf{f}^*(X)) \geq \ell_a\right]$ for lengths equal to

$$\ell_a = \lceil a + \log_2(2^{-a} + 2^{-a} M(2^a)) \rceil, \tag{85}$$

where the second term is negative and represents the exact gain with respect to the information spectrum of the source.

For later use we observe that if we let $M_X^+(\beta)$ be the number of masses with probability larger or equal than $\frac{1}{\beta}$, then,[8]

$$\begin{aligned} M_X^+(\beta) &= \sum_{x \in \mathcal{X}} 1\left\{P_X(x) \geq \frac{1}{\beta}\right\} && (86)\\ &= \mathbb{E}\left[\exp\left(\imath_X(X)\right) 1\left\{\imath_X(X) \leq \log_2 \beta\right\}\right]. && (87) \end{aligned}$$

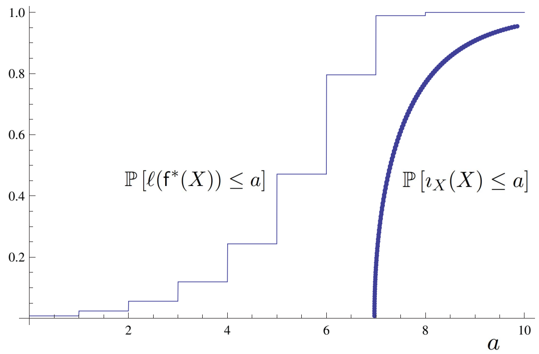

Fig. 1. Cumulative distribution functions of $\ell(\mathsf{f}^*(X))$ and $\imath_X(X)$ when X is the number of tails obtained in 10,000 fair coin flips.

Figure 1 shows the cumulative distribution functions of $\ell(\mathsf{f}^*(X))$ and $\imath_X(X)$ when X is a binomially distributed random variable: the number of tails obtained in 10,000 fair coin flips. Therefore, $\imath_X(X)$ ranges from $10^4 - \log_2 \binom{10^4}{5000} \approx 6.97$ to 10^4 and,

$$H(X) = 7.69 \tag{88}$$

$$\mathbb{E}[\ell(\mathsf{f}^*(X))] = 6.29, \tag{89}$$

where all figures are in bits. This example illustrates that in the non-asymptotic regime, the optimum coding length may be substantially smaller than the information (cf. Footnote 6).

D. Exact Behavior of Random Binning

The following result gives an exact expression for the performance of random binning for arbitrary sources (cf. [32] for the exact performance of random coding in channel coding), as a function of the cumulative distribution function of the random variable $\imath_X(X)$ via (72). In binning, the compressor is no longer constrained to be an injective mapping. When the label received by the decompressor can be explained by more than one source realization, it chooses the most likely one, breaking ties arbitrarily.

Theorem 8: Averaging uniformly over all binning compressors $\mathsf{f}\colon \mathcal{X} \to \{1, 2, \ldots N\}$, results in an expected error probability equal to

$$1 - \mathbb{E}\left[\Psi(X) \sum_{\ell=0}^{J(X)-1} \frac{\binom{J(X)-1}{\ell}}{(N-1)^\ell (1+\ell)}\right], \tag{90}$$

where $M(\cdot)$ is given in (72), the number of masses whose probability is equal to $P_X(x)$ is denoted by:

$$J(x) = \frac{\mathbb{P}[P_X(X) = P_X(x)]}{P_X(x)}, \tag{91}$$

and

$$\Psi(x) = \left(1 - \frac{1}{N}\right)^{M(\frac{1}{P_X(x)}) + J(x) - 1} \tag{92}$$

Proof: For the purposes of the proof, it is convenient to assume that ties are broken uniformly at random among the most likely source outcomes in the bin. To verify (90), note that given that the source realization is x_0:

1) The number of masses with probability strictly higher than that of x_0 is $M(\frac{1}{P_X(x_0)})$.

[8] Where typographically convenient we use $\exp(a) = 2^a$.

2) Correct decompression of x_0 requires that no x with $P_X(x) > P_X(x_0)$ be assigned to the same bin as x_0. This occurs with probability:

$$\left(1 - \frac{1}{N}\right)^{M(\frac{1}{P_X(x_0)})}. \tag{93}$$

3) If, in addition to x_0, its bin contains ℓ masses with the same probability as x_0, correct decompression occurs, assuming 2) is satisfied, with probability $\frac{1}{1+\ell}$.

4) The probability that there are ℓ masses with the same probability as x_0 in the same bin is equal to:

$$\binom{J(x_0)-1}{\ell}\left(1 - \frac{1}{N}\right)^{J(x_0)-\ell-1}\frac{1}{N^{\ell}}. \tag{94}$$

Then, (90) follows since all the bin assignments are independent. ∎

Theorem 8 leads to an achievability bound for both almost-lossless fixed-to-fixed compression and lossless fixed-to-variable compression. However, in view of the simplicity and tightness of Theorem 3, the main usefulness of Theorem 8 is to gauge the suboptimality of random binning in the finite (in fact, rather short because of computational complexity) blocklength regime.

III. Minimal Expected Length

Recall the definition of the best achievable rate $\bar{R}(n)$ in Section I, expressed in terms of f_n^* as in (10). An immediate consequence of Theorem 3 is the bound,

$$n\bar{R}(n) = \mathbb{E}\left[\ell(\mathsf{f}^*(X))\right] \le H(X), \tag{95}$$

Indeed, by lifting the prefix condition it is possible to beat the entropy on average as we saw in the asymptotic results (33) and (34). Lower bounds on the minimal average length as a function of $H(X)$ can be found in [1], [3], [5], [11], [23], [33], [38], [42]. An explicit expression can be obtained easily by labeling the outcomes as the positive integers with decreasing probabilities as in the proof of Theorem 3:

$$\begin{aligned}\mathbb{E}[\ell(\mathsf{f}^*(X))] &= \mathbb{E}[\lfloor \log_2 X \rfloor] && (96)\\ &= \sum_{k=1}^{\infty} \mathbb{P}[\lfloor \log_2 X\rfloor \ge k] && (97)\\ &= \sum_{k=1}^{\infty} \mathbb{P}[X \ge 2^k]. && (98)\end{aligned}$$

In contrast, the corresponding minimum length for prefix codes can be computed numerically, in the special case of finite alphabets, using the Huffman algorithm, but no exact expression in terms of P_X is known.

Example 2: The average number of bits required to encode at which flip of a fair coin the first tail appears is equal to

$$\begin{aligned}\sum_{k=1}^{\infty} \mathbb{P}[X \ge 2^k] &= \sum_{k=1}^{\infty}\sum_{j=2^k}^{\infty} 2^{-j} && (99)\\ &= 2\sum_{k=1}^{\infty} 2^{-2^k} && (100)\\ &\approx 0.632843, && (101)\end{aligned}$$

since, in this case, X is a geometric random variable with $\mathbb{P}[X = j] = 2^{-j}$. In contrast, imposing a prefix constraint disables any compression: The optimal prefix code consists of all, possibly empty, strings of `0`s terminated by `1`, achieving an average length of 2.

Example 3: Suppose X_M is equiprobable on a set of M elements:

1) The minimal average length is

$$\begin{aligned}&\mathbb{E}\left[\ell(\mathsf{f}^*(X_M))\right]\\ &= \lfloor \log_2 M\rfloor + \frac{1}{M}\left(2 + \lfloor \log_2 M\rfloor - 2^{\lfloor \log_2 M\rfloor + 1}\right) && (102)\end{aligned}$$

which simplifies to

$$\mathbb{E}\left[\ell(\mathsf{f}^*(X_M))\right] = \frac{(M+1)\log_2(M+1)}{M} - 2, \tag{103}$$

when $M+1$ is a power of 2.

2) For large alphabet sizes M,

$$\begin{aligned}&\limsup_{M\to\infty}\left\{H(X_M) - \mathbb{E}\left[\ell(\mathsf{f}^*(X_M))\right]\right\} = 2 && (104)\\ &\liminf_{M\to\infty}\left\{H(X_M) - \mathbb{E}\left[\ell(\mathsf{f}^*(X_M))\right]\right\}\\ &= 1 + \log_2 e - \log_2\log_2 e, && (105)\end{aligned}$$

where the entropy is expressed in bits.

Theorem 9: For any source $\mathbf{X}$ with finite entropy rate,

$$H(\mathbf{X}) = \limsup_{n\to\infty}\frac{1}{n}H(X^n) < \infty, \tag{106}$$

the normalized minimal average length satisfies:

$$\limsup_{n\to\infty}\bar{R}(n) = H(\mathbf{X}). \tag{107}$$

Proof: The achievability (upper) bound in (107) holds in view of (95). In the reverse direction, we invoke the bound [1]:

$$H(X^n) - \mathbb{E}[\ell(\mathsf{f}_n^*(X^n))] \le \log_2(H(X^n)+1) + \log_2 e. \tag{108}$$

Upon dividing both sides of (108) by n and taking lim sup the desired result follows, since for any $\delta > 0$, for all sufficiently large n, $H(X^n) \le nH(\mathbf{X}) + n\delta$. ∎

In view of (40), we see that the penalty incurred on the *average rate* by the prefix condition vanishes asymptotically in the very wide generality allowed by Theorem 9. In fact, the same proof we used for Theorem 9 shows the following result:

Theorem 10: For any (not necessarily serial, i.e. $\mathcal{A}_n$ is not necessarily a Cartesian product) source $\mathbf{X} = \{P_{X^{(n)}} \in \mathcal{A}_n\}$,

$$\lim_{n\to\infty}\frac{\bar{R}(n)}{\frac{1}{n}H(X^{(n)})} = \lim_{n\to\infty}\frac{\mathbb{E}[\ell(\mathsf{f}_n^*(X^{(n)}))]}{H(X^{(n)})} = 1, \tag{109}$$

as long as $H(X^{(n)})$ diverges.

IV. Pointwise Asymptotics

A. Normalized Pointwise Redundancy

Before turning to the precise evaluation of the best achievable rate $R^*(n,\epsilon)$, in this section we examine the asymptotic behavior of the actual codelengths $\ell(\mathsf{f}_n(X^n))$ of arbitrary compressors f_n. More specifically, we examine the difference between the codelength $\ell(\mathsf{f}_n(X^n))$ and the information function, a quantity sometimes called the "pointwise redundancy."

Theorem 11: For any finite-alphabet discrete source $\mathbf{X}$ and any positive divergent deterministic sequence κ_n such that

$$\lim_{n\to\infty} \frac{\log n}{\kappa_n} = 0, \tag{110}$$

the following hold.

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

$$\liminf_{n\to\infty} \frac{1}{\kappa_n}\left(\ell(\mathsf{f}_n(X^n)) - \imath_{X^n}(X^n)\right) \geq 0. \tag{111}$$

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

$$\lim_{n\to\infty} \frac{1}{\kappa_n}\left(\ell(\mathsf{f}_n^*(X^n)) - \imath_{X^n}(X^n)\right) = 0, \tag{112}$$

Proof: Part (a): We invoke the general converse in Theorem 5, with X^n and $\mathcal{A}^n$ in place of X and $\mathcal{X}$, respectively. Fixing arbitrary $\epsilon > 0$ and letting $\tau = \tau_n = \epsilon\kappa_n$, we obtain that

$$\begin{aligned}\mathbb{P}\big[\ell(\mathsf{f}_n(X^n)) &\leq \imath_{X^n}(X^n) - \epsilon\kappa_n\big] \\ &\leq 2^{\log_2 n - \epsilon\kappa_n}\left(\log_2|\mathcal{A}| + 1\right),\end{aligned} \tag{113}$$

which is summable in n. Therefore, the Borel-Cantelli lemma implies that the lim sup of the event on the left side of (113) has zero probability, or equivalently, with probability 1,

$$\ell(\mathsf{f}_n(X^n)) - \imath_{X^n}(X^n) \geq -\epsilon\kappa_n \tag{114}$$

is violated only a finite number of times. Since ϵ can be chosen arbitrarily small, (111) follows. Part (b) follows from (a) and (47). ■

B. Stationary Ergodic Sources

Theorem 9 states that for any discrete process $\mathbf{X}$ the expected rate of the optimal codes f_n^* satisfy,

$$\limsup_{n\to\infty} \frac{1}{n}\mathbb{E}[\ell(\mathsf{f}_n^*(X^n))] = H(\mathbf{X}). \tag{115}$$

The next result shows that if the source is stationary and ergodic, then the same asymptotic relation holds not just in expectation, but also with probability 1. Moreover, no compressor can beat the entropy rate asymptotically with positive probability. The corresponding results for prefix codes were established in [2], [18], [19].

Theorem 12: Suppose that $\mathbf{X}$ is a stationary ergodic source with entropy rate $H(\mathbf{X})$.

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

$$\liminf_{n\to\infty} \frac{1}{n}\ell(\mathsf{f}_n(X^n)) \geq H(\mathbf{X}), \quad \text{w.p.1.} \tag{116}$$

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

$$\lim_{n\to\infty} \frac{1}{n}\ell(\mathsf{f}_n^*(X^n)) = H(\mathbf{X}), \quad \text{w.p.1.} \tag{117}$$

Proof: The Shannon-Macmillan-Breiman theorem states that

$$\frac{1}{n}\imath_{X^n}(X^n) \to H(\mathbf{X}), \quad \text{w.p.1.} \tag{118}$$

Therefore, the result is an immediate consequence of Theorem 11 with $\kappa_n = n$. ■

C. Stationary Ergodic Markov Sources

We assume that the source is a stationary ergodic (first-order) Markov chain, with transition kernel,

$$P_{X'|X}(x'\,|\,x) \quad (x, x') \in \mathcal{A}^2, \tag{119}$$

on the finite alphabet $\mathcal{A}$. Further restricting the source to be Markov enables us to analyze more precisely the behavior of the information random variables and, in particular, the zero-mean random variables,

$$Z_n = \frac{\imath_{X^n}(X^n) - H(X^n)}{\sqrt{n}}, \tag{120}$$

will be seen to be asymptotically normal with variance given by the varentropy rate. This generalizes the varentropy of a single random variable defined in (32).

Definition 1: The *varentropy rate* of a random process $\mathbf{X}$ is:

$$\sigma^2(\mathbf{X}) = \limsup_{n\to\infty} \frac{1}{n}\mathsf{Var}(\imath_{X^n}(X^n)). \tag{121}$$

Remarks:

1) If $\mathbf{X}$ is a stationary memoryless process each of whose letters X_n is distributed according to P_X, then the varentropy rate of $\mathbf{X}$ is equal to the varentropy of X. The varentropy of X is zero if and only if it is equiprobable on its support.
2) In contrast to the first moment, we do not know whether stationarity is sufficient for lim sup = lim inf in (121). The difficulty appears to be that, unlike $\imath_{X^n}$, $\imath^2_{X^n}$ does not satisfy a chain rule.
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 the distribution $P_{X'|X}(\cdot\,|\,a)$, then the varentropy of the chain is, in general, not given by $\mathbb{E}[\sigma^2(X_0)]$. The reason is the nonzero correlation between the random variables $\{\imath_{X_k|X_{k-1}}(X_k|X_{k-1})\}$.
4) The varentropy rate of Markov sources is typically nonzero. For example, for a first order Markov chain it was observed in [20], [44] that $\sigma^2 = 0$ 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 that $X_1 = x_1$) q^n, for some constant $0 \leq q \leq 1$.

The simplest nontrivial example of the varentropy of a source with memory is the following.

Example 4: Consider a binary symmetric homogeneous stationary Markov chain with

$$\mathbb{P}[X_2 = 0|X_1 = 1] = \mathbb{P}[X_2 = 1|X_1 = 0] = \alpha \tag{122}$$

The entropy in bits is

$$\begin{aligned} H(X^n) &= H(X_1) + (n-1)H(X_2|X_1) &(123)\\ &= 1 + (n-1)h(\alpha) &(124)\end{aligned}$$

Furthermore, it is easy to check that for $k = 1, 2, ...$

$$\begin{aligned}&\mathbb{E}[\log^2 P_{X_{k+1}|X_k}(X_{k+1}|X_k)] \\ &\quad= (1-\alpha)\log^2(1-\alpha) + \alpha\log^2\alpha\end{aligned} \tag{125}$$

while for $k = 2, 3, \dots$

$$\mathbb{E}[\log P_{X_{k+1}|X_k}(X_{k+1}|X_k) \log P_{X_2|X_1}(X_2|X_1)] = h^2(\alpha) \quad (126)$$

We can then write

$$\begin{aligned}
&\mathbb{E}\left[\log^2 P_{X^n}(X^n)\right] \\
&= \mathbb{E}\left[\left(1 - \sum_{k=1}^{n-1} \log P_{X_{k+1}|X_k}(X_{k+1}|X_k)\right)^2\right] && (127) \\
&= 1 + 2(n-1)h(\alpha) + (n-1)(n-2)h^2(\alpha) \\
&\quad +(n-1)\mathbb{E}[\log^2 P_{X_2|X_1}(X_2|X_1)] && (128) \\
&= H^2(X^n) + (n-1)\alpha(1-\alpha)\log^2\frac{1-\alpha}{\alpha} && (129)
\end{aligned}$$

Therefore, we conclude that

$$\mathsf{Var}(\imath_{X^n}(X^n)) = (n-1)\alpha(1-\alpha)\log^2\frac{1-\alpha}{\alpha} \quad (130)$$

and consequently

$$\sigma^2(\mathbf{X}) = \alpha(1-\alpha)\log^2\frac{1-\alpha}{\alpha} \quad (131)$$

which coincides with the varentropy of the Bernoulli source with bias α.

The following result is known (see [31] and the references therein.) We include its proof for completeness and for later reference.

Theorem 13: Let $\mathbf{X}$ be a stationary ergodic finite-state Markov chain.

(i) The varentropy rate $\sigma^2(\mathbf{X})$ is also equal to the corresponding lim inf of the normalized variances in (121), and it is finite.

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

$$\frac{\imath_{X^n}(X^n) - H(X^n)}{\sqrt{n}} \longrightarrow N(0, \sigma^2(\mathbf{X})), \quad (132)$$

in distribution.

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

$$\limsup_{n\to\infty} \frac{\imath_{X^n}(X^n) - H(X^n)}{\sqrt{2n \ln\ln n}} = \sigma(\mathbf{X}), \quad \text{w.p.1} \quad (133)$$

$$\liminf_{n\to\infty} \frac{\imath_{X^n}(X^n) - H(X^n)}{\sqrt{2n \ln\ln n}} = -\sigma(\mathbf{X}), \quad \text{w.p.1} \quad (134)$$

Proof: (i) and (ii): Consider the bivariate Markov chain $\{\tilde{X}_n = (X_n, X_{n+1})\}$ on the alphabet $\mathcal{B} = \{(x, y) \in \mathcal{A}^2 : P_{X'|X}(y|x) > 0\}$ and the function $f : \mathcal{B} \to \mathbb{R}$ defined by $f(x, y) = \imath_{X'|X}(y|x)$. Since $\mathbf{X}$ is stationary and ergodic, so is $\{\tilde{X}_n\}$, hence, by the central limit theorem for functions of Markov chains [6],

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\left(\imath_{X^n|X_1}(X^n|X_1) - H(X^n|X_1)\right) \\
&\quad = \frac{1}{\sqrt{n}}\sum_{i=1}^{n-1}(f(\tilde{X}_i) - \mathbb{E}[f(\tilde{X}_i)]), && (135)
\end{aligned}$$

converges in distribution to the zero-mean Gaussian law with finite variance

$$\Sigma^2 = \lim_{n\to\infty} \frac{1}{n}\mathsf{Var}(\imath_{X^n|X_1}(X^n|X_1)). \quad (136)$$

Furthermore, since

$$\begin{aligned}
\imath_{X^n}(X^n) - H(X^n) &= \imath_{X^n|X_1}(X^n|X_1) - H(X^n|X_1) \\
&\quad + \left(\imath_{X_1}(X_1) - H(X_1)\right), && (137)
\end{aligned}$$

where the second term is bounded, (132) must hold and we must have $\Sigma^2 = \sigma^2(\mathbf{X})$.

(iii): Normalizing (135) by $\sqrt{2n \ln\ln n}$ instead of $\sqrt{n}$, we can invoke the law of the iterated logarithm for functions of Markov chains [6] to show that the lim sup / lim inf of the sum behave as claimed. Since, upon normalization, the last term in (137) vanishes almost surely, $\imath_{X^n}(X^n) - H(X^n)$ must exhibit the same asymptotic behavior. ■

Similarly, the following result readily follows from Theorem 13 and Theorem 11 with $\kappa_n = \sqrt{2n \ln\ln n}$.

Theorem 14: Suppose $\mathbf{X}$ is a stationary ergodic Markov chain with entropy rate $H(\mathbf{X})$ and varentropy rate $\sigma^2(\mathbf{X})$. Then:

(i)

$$\frac{\ell(\mathsf{f}_n^*(X^n)) - nH(\mathbf{X})}{\sqrt{n}} \longrightarrow N(0, \sigma^2(\mathbf{X})), \quad (138)$$

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

$$\limsup_{n\to\infty} \frac{\ell(\mathsf{f}_n(X^n)) - H(X^n)}{\sqrt{2n \ln\ln n}} \ge \sigma(\mathbf{X}), \quad \text{w.p.1;} \quad (139)$$

$$\liminf_{n\to\infty} \frac{\ell(\mathsf{f}_n(X^n)) - H(X^n)}{\sqrt{2n \ln\ln n}} \ge -\sigma(\mathbf{X}), \quad \text{w.p.1.} \quad (140)$$

(ii) The sequence of optimal codes $\{\mathsf{f}_n^*\}$ achieves the bounds in (139) and (140) 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 than that of the optimal codes without prefix constraints.

D. Beyond Markov

The Markov sufficient condition in Theorem 13 enabled the application of the central limit theorem and the law of the iterated logarithm to the sum in (135). According to Theorem 9.1 of [31] a more general sufficient condition is that $\mathbf{X}$ be a stationary process with $\alpha(d) = O(d^{-336})$ and $\gamma(d) = O(d^{-48})$, where the mixing coefficients $\alpha(d)$ and $\gamma(d)$ are defined as:

$$\gamma(d) = \max_{a\in\mathcal{A}} \mathbb{E}\left|\imath_{X_0|X_{-\infty}^{-1}}(a|X_{-\infty}^{-1}) - \imath_{X_0|X_{-d}^{-1}}(a|X_{-d}^{-1})\right| \quad (141)$$

$$\alpha(d) = \sup\{|\mathbb{P}(B \cap A) - \mathbb{P}(B)\mathbb{P}(A)|\} \quad (142)$$

where the sup is over $A \in \mathcal{F}_{-\infty}^0$ and $B \in \mathcal{F}_d^\infty$.

Here $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_d^\infty$ denote the σ-algebras generated by the collections of random variables $(X_0, X_{-1}, \dots)$ and $(X_d, X_{d+1}, \dots)$, respectively. The $\alpha(d)$ are the *strong mixing* coefficients [4] of $\mathbf{X}$, and the $\gamma(d)$ were introduced by Ibragimov in [16]. Although these mixing conditions may be hard to verify in practice, they are fairly weak in that they require only polynomial decay of $\alpha(d)$ and $\gamma(d)$. In particular, any ergodic Markov chain of any order satisfies these conditions.

From the Lindeberg-Feller theorem [7], another sufficient condition for (132) to hold consists in assuming that the lim sup in (121) is equal to the lim inf and is finite, and that for all $\eta > 0$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\left[Z_j 1\{Z_j > \eta V(X^n)\}\right] = 0 \qquad (143)$$

where

$$Z_j = \left(\imath_{X_j|X^{j-1}}(X_j|X^{j-1}) - H(X_j|X^{j-1})\right)^2. \qquad (144)$$

V. Gaussian Approximation for Memoryless Sources

We turn our attention to the non-asymptotic behavior of the best rate $R^*(n,\epsilon)$ that can be achieved when compressing a stationary memoryless source $\mathbf{X}$ with values in the finite alphabet $\mathcal{A}$ and marginal distribution P_X, whose entropy and varentropy are denoted by $H(X)$ and $\sigma^2(X)$, respectively.

Specifically, we will derive explicit upper and lower bounds on $R^*(n,\epsilon)$ in terms of the first three moments of the information random variable $\imath_X(X)$. Although using Theorem 7 it is possible, in principle, to compute $R^*(n,\epsilon)$ exactly, it is more desirable to derive approximations that are both easier to compute and offer more intuition into the behavior of the fundamental limit $R^*(n,\epsilon)$.

Theorems 17 and 18 imply that, for all $\epsilon \in (0,1/2)$, the best achievable rate $R^*(n,\epsilon)$ satisfies,

$$\frac{c}{n} \le R^*(n,\epsilon) - \left[H(X) + \frac{\sigma(X)}{\sqrt{n}}Q^{-1}(\epsilon) - \frac{\log_2 n}{2n}\right] \le \frac{c'}{n}. \qquad (145)$$

The upper bound is valid for all n, and the lower bound is valid for $n \ge n_0$. Explicit values are derived for the constants n_0, c and c'. In view of Theorem 1 which states that minimizing the probability that the encoded length exceeds a given bound, the prefix constraint only incurs one bit penalty, essentially the same results as in (145) hold for prefix codes:

$$\frac{c}{n} \le R_{\mathsf{p}}(n,\epsilon) - \left[H(X) + \frac{\sigma(X)}{\sqrt{n}}Q^{-1}(\epsilon) - \frac{\log_2 n}{2n}\right] \le \frac{c'+1}{n}. \qquad (146)$$

The bounds in (145) and (146) result in the Gaussian approximation (44) stated in Section I.

Before establishing the precise non-asymptotic relations leading to (145) and (146), we illustrate their utility. To facilitate this, note that Theorem 3 immediately yields the following simple bound:

Theorem 15: For all $n \ge 1$, $\epsilon > 0$,

$$R^*(n,\epsilon) \le R^{\mathsf{u}}(n,\epsilon), \qquad (147)$$

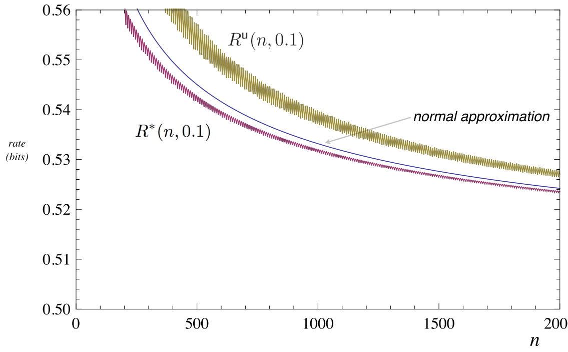

Fig. 2. The optimum rate $R^*(n,0.1)$, the Gaussian approximation $\tilde{R}^*(n,0.1)$ in (149), and the upper bound $R^{\mathsf{u}}(n,0.1)$, for a Bernoulli-0.11 source and blocklengths $200 \le n \le 2000$.

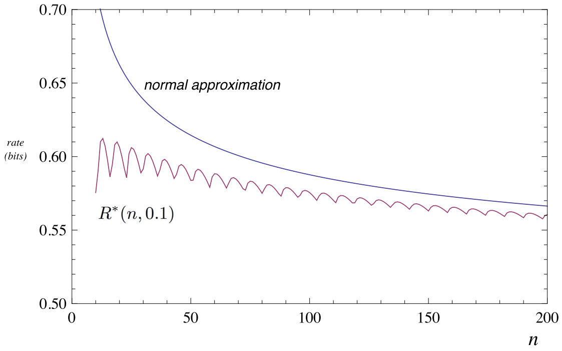

Fig. 3. The optimum rate $R^*(n,0.1)$ and the Gaussian approximation $\tilde{R}^*(n,0.1)$ in (149), for a Bernoulli-0.11 source and blocklengths $10 \le n \le 200$.

where $R^{\mathsf{u}}(n,\epsilon)$ is the quantile function of the information spectrum, i.e., the lowest R such that:

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\imath_X(X_i) \ge R\right] \le \epsilon. \qquad (148)$$

In Figure 2, we exhibit the behavior of the fundamental compression limit $R^*(n,\epsilon)$ in the case of coin flips with bias $0.11 = h^{-1}(0.5)$ (for which $H(X) = 0.5$ bits). In particular, we compare $R^*(n,\epsilon)$ and $R^{\mathsf{u}}(n,\epsilon)$ for $\epsilon = 0.1$. The non-monotonic nature of both $R^*(n,\epsilon)$ and $R^{\mathsf{u}}(n,\epsilon)$ with n is not surprising: Although, the larger the value of n, the less we are at the mercy of the source randomness, we also need to compress more information. Figure 2 also illustrates that $R^*(n,\epsilon)$ is tracked rather closely by the Gaussian approximation,

$$\tilde{R}^*(n,\epsilon) = H(X) + Q^{-1}(\epsilon)\frac{\sigma(X)}{\sqrt{n}} - \frac{1}{2n}\log_2 n, \qquad (149)$$

suggested by (145).

Figure 3 focuses the comparison between $R^*(n,0.1)$ and $\tilde{R}^*(n,0.1)$ on the short blocklength range up to 200 not shown in Figure 2. For $n > 60$, the discrepancy between the two never exceeds 4%.

The remainder of the section is devoted to justifying the use of (149) as an accurate approximation to $R^*(n,\epsilon)$. To that end, in Theorems 18 and 17 we establish the bounds given in (145). Their derivation requires that we overcome two technical hurdles:

1) The distribution function of the optimal encoding length is not the same as the distribution of $\frac{1}{n}\sum_{i=1}^{n} \imath_X(X_i)$;
2) The distribution of $\frac{1}{n}\sum_{i=1}^{n} \imath_X(X_i)$ is only approximately Gaussian.

To cope with the second hurdle we will appeal to the classical Berry-Esséen bound [22], [30]:

Theorem 16: Let $\{Z_i\}$ be independent and identically distributed random variables with zero mean and unit variance, and let $\bar{Z}$ be standard normal. Then, for all $n \geq 1$ and any a:

$$\left|\mathbb{P}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i \leq a\right] - \mathbb{P}\left[\bar{Z} \leq a\right]\right| \leq \frac{\mathbb{E}[|Z_1|^3]}{2\sqrt{n}}. \tag{150}$$

Invoking the Berry-Esséen bound, Strassen [37] claimed the following approximation for $n > \frac{19600}{\delta^{16}}$,

$$\left|R^*(n,\epsilon) - \tilde{R}^*(n,\epsilon)\right| \leq \frac{140}{\delta^8}, \tag{151}$$

where,

$$\delta \leq \min\left\{\sigma(X), \epsilon, 1-\epsilon, \mu_3^{-1/3}\right\} \tag{152}$$

$$\mu_3 = \mathbb{E}[|\imath_X(X) - H(X)|^3]. \tag{153}$$

Unfortunately, there appears to be a gap in how Strassen [37] justifies the application of an asymptotic form of the Berry-Esséen theorem with uniform convergence, in order to bound the error introduced by taking expectations with respect to Gaussian approximations (cf. equations (2.17), (3.18) and the displayed equation between (3.15) and (3.16) in [37]).

The following achievability result holds for all blocklengths.

Theorem 17: For all $0 < \epsilon \leq \frac{1}{2}$ and all $n \geq 1$,

$$\begin{aligned} R^*(n,\epsilon) \leq\ & H(X) + \frac{\sigma(X)}{\sqrt{n}} Q^{-1}(\epsilon) - \frac{\log_2 n}{2n} \\ & + \frac{1}{n}\log_2\left(\frac{\log_2 e}{\sqrt{2\pi\sigma^2(X)}} + \frac{\mu_3}{\sigma^3(X)}\right) \\ & + \frac{1}{n}\frac{\mu_3}{\sigma^2(X)\phi(\Phi^{-1}(\Phi(Q^{-1}(\epsilon)) + \frac{\mu_3}{\sigma^3(X)\sqrt{n}}))}, \end{aligned} \tag{154}$$

as long as the varentropy $\sigma^2(X)$ is strictly positive, where $\Phi = 1 - Q$ and $\phi = \Phi'$ are the standard Gaussian distribution function and density, respectively, and μ_3 is the third absolute moment of $\imath_X(X)$ defined in (153).

Proof: The proof follows Strassen's construction, but the essential approximation steps are different. The positive constant β_n is uniquely defined by:

$$\mathbb{P}\left[\imath_{X^n}(X^n) \leq \log_2 \beta_n\right] \geq 1 - \epsilon, \tag{155}$$

$$\mathbb{P}\left[\imath_{X^n}(X^n) < \log_2 \beta_n\right] < 1 - \epsilon. \tag{156}$$

Since the information spectrum (i.e., the distribution function of the information random variable $\imath_{X^n}(X^n)$) is piecewise constant, $\log_2 \beta_n$ is the location of the jump where the information spectrum reaches (or exceeds for the first time) the value $1-\epsilon$. Furthermore, defining the normalized constant,

$$\lambda_n = \frac{\log_2 \beta_n - nH(X)}{\sqrt{n}\sigma(X)}, \tag{157}$$

the probability in the left side of (155) satisfies

$$\begin{aligned} & \mathbb{P}\left[\frac{\imath_{X^n}(X^n) - nH(X)}{\sqrt{n}\sigma(X)} \leq \lambda_n\right] \\ & \qquad \leq \Phi(\lambda_n) + \frac{\mu_3}{2\sigma^3(X)\sqrt{n}}, \end{aligned} \tag{158}$$

where we have applied Theorem 16. Analogously ((150) also holds for $\mathbb{P}[\frac{1}{\sqrt{n}}\sum Z_i < a]$), we obtain,

$$\begin{aligned} & \mathbb{P}\left[\frac{\imath_{X^n}(X^n) - nH(X)}{\sqrt{n}\sigma(X)} < \lambda_n\right] \\ & \qquad \geq \Phi(\lambda_n) - \frac{\mu_3}{2\sigma^3(X)\sqrt{n}}. \end{aligned} \tag{159}$$

Since $1 - \epsilon$ is sandwiched between the right sides of (158) and (159), as $n \to \infty$ we must have $\lambda_n \to \lambda$, where,

$$\lambda = \Phi^{-1}(1-\epsilon) = Q^{-1}(\epsilon). \tag{160}$$

By a simple first-order Taylor bound,

$$\lambda_n \leq \Phi^{-1}\left(\Phi(\lambda) + \frac{\mu_3}{2\sigma^3(X)\sqrt{n}}\right) \tag{161}$$

$$= \lambda + \frac{\mu_3}{2\sigma^3(X)\sqrt{n}}(\Phi^{-1})'(\xi_n) \tag{162}$$

$$= \lambda + \frac{\mu_3}{2\sigma^3(X)\sqrt{n}}\frac{1}{\phi(\Phi^{-1}(\xi_n))}, \tag{163}$$

for some $\xi_n \in [\Phi(\lambda), \Phi(\lambda) + \frac{\mu_3}{2\sigma^3(X)\sqrt{n}}]$. Since $\epsilon \leq 1/2$, we have $\lambda \geq 0$ and $\Phi(\lambda) \geq 1/2$, so that $\xi_n \geq 1/2$. And since $\Phi^{-1}(t)$ is strictly increasing for all t, while ϕ is strictly decreasing for $t \geq 0$, from (163) we obtain,

$$\lambda_n \leq \lambda + \frac{\mu_3}{2\sigma^3(X)\sqrt{n}}\frac{1}{\phi(\Phi^{-1}(\Phi(\lambda) + \frac{\mu_3}{2\sigma^3(X)\sqrt{n}}))}. \tag{164}$$

The event E_n in the left side of (155) contains all the "high probability strings," and has probability $\geq 1-\epsilon$. Its cardinality is $M_X^+(\beta_n)$, defined in (86) (with $X \leftarrow X^n$). Therefore, denoting,

$$\varphi(t) = 2^{-t}1\{t \geq 0\} \tag{165}$$

$$Y_i = \frac{1}{\sigma(X)}\left(\imath_X(X_i) - H(X)\right), \tag{166}$$

we obtain,

$$R^*(n,\epsilon) \leq \frac{1}{n}\log_2 M_X^+(\beta_n) \tag{167}$$

$$\begin{aligned} = \frac{1}{n}\log_2 \mathbb{E}\Big[& \exp\left(\imath_{X^n}(X^n)\right) \\ & \times 1\left\{\imath_{X^n}(X^n) \leq \log_2 \beta_n\right\}\Big] \end{aligned} \tag{168}$$

$$= H(X) + \lambda_n \frac{\sigma(X)}{\sqrt{n}} + \frac{1}{n}\log_2 \alpha_n, \tag{169}$$

with

$$\alpha_n = \mathbb{E}\left[\varphi(\log_2 \beta_n - \imath_{X^n}(X^n))\right] \tag{170}$$

$$= \mathbb{E}\left[\varphi\left(\sqrt{n}\sigma(X)\left(\lambda_n - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i\right)\right)\right], \tag{171}$$

where we have used (157), $\exp(a) = 2^a$, and $\{Y_i\}$ are independent, identically distributed, with zero mean and unit variance.

Let $\bar{\alpha}_n$ be defined as (171) except that Y_i are replaced by $\bar{Y}_i$ which are standard normal. Then, straightforward algebra yields,

$$\bar{\alpha}_n = \mathbb{E}\left[2^{-\sqrt{n}\sigma(X)(\lambda_n - \bar{Y}_1)} 1\{\bar{Y}_1 \leq \lambda_n\}\right] \tag{172}$$

$$= \int_0^\infty 2^{-x} \frac{e^{-\frac{(x-\lambda_n\sigma(X)\sqrt{n})^2}{2\sigma^2(X)n}}}{\sqrt{2\pi\sigma^2(X)n}}\, dx \tag{173}$$

$$\leq \frac{\log_2 e}{\sqrt{2\pi\sigma^2(X)n}}. \tag{174}$$

To deal with the fact that the random variables in (171) are not normal, we apply the Lebesgue-Stieltjes formula for integration by parts to (171). Denoting the distribution of the normalized sum in (171) by $F_n(t)$, α_n becomes,

$$\alpha_n = \int_{-\infty}^{\lambda_n} 2^{-(\sqrt{n}\sigma(X)(\lambda_n - t))} dF_n(t) \tag{175}$$

$$= F_n(\lambda_n) - \int_{-\infty}^{\lambda_n} F_n(t)\sqrt{n}\sigma(X) 2^{-(\sqrt{n}\sigma(X)(\lambda_n - t))} dt \log_e 2 \tag{176}$$

$$= \bar{\alpha}_n + F_n(\lambda_n) - \Phi(\lambda_n) - \sqrt{n}\sigma(X)\log_e 2 \times \int_{-\infty}^{\lambda_n} (F_n(t) - \Phi(t)) 2^{-(\sqrt{n}\sigma(X)(\lambda_n - t))} dt \tag{177}$$

$$\leq \bar{\alpha}_n + \frac{\mu_3}{2\sigma^3(X)\sqrt{n}} + \frac{\mu_3}{2\sigma^2(X)} \int_{-\infty}^{\lambda_n} 2^{-(\sqrt{n}\sigma(X)(\lambda_n - t))} dt \log_e 2 \tag{178}$$

$$= \bar{\alpha}_n + \frac{\mu_3}{\sigma^3(X)\sqrt{n}} \tag{179}$$

$$\leq \frac{1}{\sqrt{n}} \left(\frac{\log_2 e}{\sqrt{2\pi\sigma^2(X)}} + \frac{\mu_3}{\sigma^3(X)} \right), \tag{180}$$

where (178) results from applying Theorem 16 twice. The desired result now follows from (169) after assembling the bounds on λ_n and α_n in (164) and (180), respectively. ■

Next we give a complementary converse result.

Theorem 18: For all $0 < \epsilon < \frac{1}{2}$ and all n such that

$$n > n_0 = \frac{1}{4}\left(1 + \frac{\mu_3}{2\sigma^3(X)}\right)^2 \frac{1}{\left(\phi(Q^{-1}(\epsilon))Q^{-1}(\epsilon)\right)^2}, \tag{181}$$

the following lower bound holds,

$$R^*(n,\epsilon) \geq H(X) + \frac{\sigma(X)}{\sqrt{n}} Q^{-1}(\epsilon) - \frac{\log_2 n}{2n} - \frac{\frac{\mu_3}{2} + \sigma^3(X)}{n\sigma^2(X)\phi(Q^{-1}(\epsilon))}, \tag{182}$$

as long as the varentropy $\sigma^2(X)$ is strictly positive.

Proof: Let

$$\eta = \frac{\frac{\mu_3}{2\sigma^2(X)} + \sigma(X)}{\phi(Q^{-1}(\epsilon))}, \tag{183}$$

and consider

$$\mathbb{P}\left[\sum_{i=1}^n \imath_X(X_i) \geq nH(X) + \sigma(X)\sqrt{n}Q^{-1}(\epsilon) - \eta\right]$$

$$= \mathbb{P}\left[\sum_{i=1}^n \frac{\imath_X(X_i) - H(X)}{\sigma(X)\sqrt{n}} \geq Q^{-1}(\epsilon) - \frac{\eta}{\sigma(X)\sqrt{n}}\right] \tag{184}$$

$$\geq Q\left(Q^{-1}(\epsilon) - \frac{\eta}{\sigma(X)\sqrt{n}}\right) - \frac{\mu_3}{2\sigma^3(X)\sqrt{n}} \tag{185}$$

$$\geq \epsilon + \frac{\eta}{\sigma(X)\sqrt{n}}\phi(Q^{-1}(\epsilon)) - \frac{\mu_3}{2\sigma^3(X)\sqrt{n}} \tag{186}$$

$$= \epsilon + \frac{1}{\sqrt{n}}, \tag{187}$$

where (185) follows from Theorem 16, and (186) follows from

$$Q(a - \Delta) \geq Q(a) + \Delta\phi(Q(a)), \tag{188}$$

which holds at least as long as

$$a > \frac{\Delta}{2} > 0. \tag{189}$$

Letting $a = Q^{-1}(\epsilon)$ and $\Delta = \frac{\eta}{\sigma(X)\sqrt{n}}$, (189) is equivalent to (181).

We proceed to invoke Theorem 4 with $X \leftarrow X^n$, k equal to n times the right side of (182), and $\tau = \frac{1}{2}\log_2 n$. In view of the definition of $R^*(n,\epsilon)$ and (184)–(187), the desired result follows. ■

VI. Gaussian Approximation for Markov Sources

Let $\mathbf{X}$ be an irreducible, aperiodic, kth order Markov chain on the finite alphabet $\mathcal{A}$, with transition probabilities,

$$P_{X'|X^k}(x_{k+1} \mid x^k), \quad x^{k+1} \in \mathcal{A}^{k+1}, \tag{190}$$

and entropy rate $H(\mathbf{X})$. Note that we do not assume that the source is stationary. In Theorem 13 of Section IV we established that the varentropy rate defined in general in equation (121), for stationary ergodic chains exists as the limit,

$$\sigma^2(\mathbf{X}) = \lim_{n\to\infty} \frac{1}{n} \mathsf{Var}(\imath_{X^n}(X^n)). \tag{191}$$

An examination of the proof shows that, by an application of the general central limit theorem for (uniformly ergodic) Markov chains [6], [28], the assumption of stationarity is not necessary, and (191) holds for all irreducible aperiodic chains.

Theorem 19: Suppose $\mathbf{X}$ is an irreducible and aperiodic kth order Markov source, and let $\epsilon \in (0, 1/2)$. Then, there is a positive constant C such that, for all n large enough,

$$nR^*(n,\epsilon) \leq nH(\mathbf{X}) + \sigma(\mathbf{X})\sqrt{n}Q^{-1}(\epsilon) + C, \tag{192}$$

where the varentropy rate $\sigma^2(\mathbf{X})$ is given by (191) and it is assumed to be strictly positive.

Theorem 20: Under the same assumptions as in Theorem 19, for all n large enough,

$$nR^*(n,\epsilon) \geq nH(\mathbf{X}) + \sigma(\mathbf{X})\sqrt{n}Q^{-1}(\epsilon) - \frac{1}{2}\log_2 n - C, \tag{193}$$

where $C > 0$ is a finite constant, possibly different from than in Theorem 19.

Remarks:

1) By definition, the lower bound in Theorem 20 also applies to $R_{\mathsf{p}}(n, \epsilon)$, while in view Theorem 1, the upper bound in Theorem 19 also applies to $R_{\mathsf{p}}(n, \epsilon)$ provided C is replaced by $C + 1$.
2) Note that, unlike the direct and converse coding theorems for memoryless sources (Theorems 17 and 18, respectively) in the results of Theorems 19 and 20 we do not give explicit bounds for the constant terms. This is because the main probabilistic tool we use in the proofs (the Berry-Esséen bound in Theorem 16) does not have an equally precise counterpart for Markov chains. Specifically, in the proof of Theorem 21 below we appeal to a Berry-Esséen bound established by Nagaev in [29], which does not give an explicit value for the multiplicative constant A in (198).
3) If we restrict our attention to the (much more narrow) class of *reversible* chains, then it is indeed possible to apply the Berry-Esséen bound of Mann [24] to obtain explicit values for the constants in Theorems 19 and 20; but the resulting values are pretty loose, drastically limiting the engineering usefulness of the resulting bounds. For example, in Mann's version of the Berry-Esséen bound, the corresponding right side of the inequality in Theorem 16 is multiplied by a factor of 13000. Therefore, we have opted for the less explicit but much more general statements given above.
4) Similar comments to those in the last two remarks apply to the observation that Theorem 19 is a weaker bound than that established in Theorem 17 for memoryless sources, by a $(1/2)\log_2 n$ term. Instead of restricting our result to the much more narrow class of reversible Markov chains, or extending the involved proof of Theorem 17 to the case of a Markovian source, we chose to illustrate how this weaker bound can be established in full generality, with a much shorter proof.
5) The proof of Theorem 19 shows that the constant in its statement can be chosen as

$$C = \frac{2A\sigma(\mathbf{X})}{\phi(Q^{-1}(\epsilon))}, \tag{194}$$

for all

$$n \geq \frac{2A^2}{\pi e(\phi(Q^{-1}(\epsilon)))^4}, \tag{195}$$

where A is the constant appearing in Theorem 21, below. Similarly, from the proof of Theorem 20 we see that the constant in its statement can be chosen as

$$C = \frac{\sigma(\mathbf{X})(A+1)}{\phi(Q^{-1}(\epsilon))} + 1, \tag{196}$$

for all,

$$n \geq \left(\frac{A+1}{Q^{-1}(\epsilon)\phi(Q^{-1}(\epsilon))}\right)^2. \tag{197}$$

Note that, in both cases, the values of the constants can easily be improved, but they still depend on the implicit constant A of Theorem 21.

As mentioned above, we will need a Berry-Esséen-type bound on the scaled information random variables,

$$\frac{\imath_{X^n}(X^n) - nH(\mathbf{X})}{\sqrt{n}\sigma(\mathbf{X})}.$$

Beyond the Shannon-McMillan-Breiman Theorem, several more refined asymptotic results have been established for this sequence; see, in particular, [16], [31], [37], [44] and the discussions in [20] and in Section IV. Unlike these asymptotic results, we will use the following non-asymptotic bound.

Theorem 21: For an ergodic, kth order Markov source $\mathbf{X}$ with entropy rate $H(\mathbf{X})$ and positive varentropy rate $\sigma^2(\mathbf{X})$, there exists a finite constant $A > 0$ such that, for all $n \geq 1$,

$$\sup_{z\in\mathbb{R}}\left|\mathbb{P}\Big[\imath_{X^n}(X^n) - nH(\mathbf{X}) > z\,\sigma(\mathbf{X})\sqrt{n}\Big] - Q(z)\right| \leq \frac{A}{\sqrt{n}}. \tag{198}$$

Proof: For integers $i \leq j$, we adopt the notation x_i^j and X_i^j for blocks of strings $(x_i, x_{i+1}, \ldots, x_j)$ and random variables $(X_i, X_{i+1}, \ldots, X_j)$, respectively. For all $x^{n+k} \in \mathcal{A}^{n+k}$ such that $P_{X^k}(x^k) > 0$ and $P_{X'|X_{j-k}^{j-1}}(x_j \,|\, x_{j-k}^{j-1}) > 0$, for $j = k+1, k+2, \ldots n+k$, we have

$$\begin{aligned}
&\imath_{X^n}(x^n) \\
&\quad= \log_2 \frac{1}{P_{X^k}(x^k)\prod_{j=k+1}^{n} P_{X'|X_{j-k}^{j-1}}(x_j \,|\, x_{j-k}^{j-1})} && (199)\\
&\quad= \sum_{j=k+1}^{k+n} \log_2 \frac{1}{P_{X'|X_{j-k}^{j-1}}(x_j \,|\, x_{j-k}^{j-1})} \\
&\qquad - \log_2 \frac{P_{X^k}(x^k)}{\prod_{j=n+1}^{n+k} P_{X'|X_{j-k}^{j-1}}(x_j \,|\, x_{j-k}^{j-1})} && (200)\\
&\quad= \sum_{j=1}^{n} f(x^{j+k}) \;+\; \Delta_n, && (201)
\end{aligned}$$

where the function $f: \mathcal{A}^{k+1} \to \mathbb{R}$ is defined on the state space

$$\mathcal{A}' = \{x^{k+1} \in \mathcal{A}^{k+1}:\;\; P_{X'|X^k}(x_{k+1}|x^k) > 0\}. \tag{202}$$

by

$$\begin{aligned}
f(x^{k+1}) &= \imath_{X'|X^k}(x_{k+1}|x^k) && (203)\\
&= \log_2 \frac{1}{P_{X'|X^k}(x_{k+1} \,|\, x^k)}, && (204)
\end{aligned}$$

and

$$\Delta_n = \log_2 \frac{\prod_{j=n+1}^{n+k} P_{X'|X_{j-k}^{j-1}}(x_j \,|\, x_{j-k}^{j-1})}{P_{X^k}(x^k)}. \tag{205}$$

Then we can bound,

$$\begin{aligned}
|\Delta_n| &\leq \delta && (206)\\
&= \max\left|\log_2\left[\frac{P_{X^k}(x^k)}{\prod_{j=n+1}^{n+k} P_{X'|X^k}(x_j \,|\, x_{j-k}^{j-1})}\right]\right| \\
&< \infty, && (207)
\end{aligned}$$

where the maximum is over the positive probability strings for which we have established (201).

Let $\{Y_n\}$ denote the first-order Markov source defined by taking overlapping $(k+1)$-blocks in the original chain,

$$Y_n = (X_n, X_{n+1}, \dots, X_{n+k}). \tag{208}$$

Since $\mathbf{X}$ is irreducible and aperiodic, so is $\{Y_n\}$. Now, since the chain $\{Y_n\}$ is irreducible and aperiodic on a finite state space, condition (0.2) of [29] is satisfied, and since the function f is bounded, Theorem 1 of [29] implies that there exists a finite constant A_1 such that, for all n,

$$\sup_{z\in\mathbb{R}}\left|\mathbb{P}\Big[\frac{\sum_{j=1}^n f(Y_j) - nH(\mathbf{X})}{\sigma(\mathbf{X})\sqrt{n}} > z\Big] - Q(z)\right| \le \frac{A_1}{\sqrt{n}}, \tag{209}$$

where the entropy rate is $H(\mathbf{X}) = \mathbb{E}[f(\tilde{Y}_1)]$ and

$$\sigma^2(\mathbf{X}) = \lim_{n\to\infty}\frac{1}{n}\mathbb{E}\Big[\Big(\sum_{j=1}^n (f(\tilde{Y}_j) - nH(\mathbf{X}))\Big)^2\Big], \tag{210}$$

where $\{\tilde{Y}_n\}$ is a stationary version of $\{Y_n\}$, that is, it has the same transition probabilities but its initial distribution is its unique invariant distribution,

$$\mathbb{P}[\tilde{Y}_1 = x^{k+1}] = \pi(x^k)P_{X'|X^k}(x_{k+1}\,|\,x^k), \tag{211}$$

where π is the unique invariant distribution of the original chain $\mathbf{X}$. Since the function f is bounded and the distribution of the chain $\{Y_n\}$ converges to stationarity exponentially fast, it is easy to see that (210) coincides with the source varentropy rate.

Let $F_n(z)$, $G_n(z)$ denote the complementary cumulative distribution functions

$$F_n(z) = \mathbb{P}\left[\imath_{X^n}(X^n) - nH(\mathbf{X}) > z\sqrt{n}\sigma(\mathbf{X})\right], \tag{212}$$

$$G_n(z) = \mathbb{P}\left[\sum_{j=1}^n f(Y_j) - nH(\mathbf{X}) > z\sqrt{n}\sigma(\mathbf{X})\right]. \tag{213}$$

Since $F_n(z)$ and $G_n(z)$ are non-increasing, (201) and (207) imply that

$$\begin{aligned} F_n(z) &\ge G_n\left(z + \frac{\delta}{\sigma\sqrt{n}}\right) && (214)\\ &\ge Q\left(z + \frac{\delta}{\sigma\sqrt{n}}\right) - \frac{A_1}{\sqrt{n}}, && (215)\\ &\ge Q(z) - \frac{A}{\sqrt{n}}, && (216)\end{aligned}$$

uniformly in z, where (215) follows from (209), and (216) holds with $A = A_1 + \delta/\sqrt{2\pi}$ since $Q'(z) = -\phi(z)$ is bounded by $-1/\sqrt{2\pi}$. A similar argument shows

$$\begin{aligned} F_n(z) &\le G_n(z - \delta/\sqrt{n}) && (217)\\ &\le Q(z - \delta/\sqrt{n}) + \frac{A_1}{\sqrt{n}} && (218)\\ &\le Q(z) + \frac{A}{\sqrt{n}}. && (219)\end{aligned}$$

Since both (216) and (219) hold uniformly in $z \in \mathbb{R}$, together they form the statement of the theorem. ∎

Proof of Theorem 19: Let

$$K_n = nH(\mathbf{X}) + \sigma\sqrt{n}Q^{-1}(\epsilon) + C, \tag{220}$$

where we abbreviate $\sigma = \sigma(\mathbf{X})$ and, for now C is an arbitrary positive constant. Theorem 3 with X^n in place of X states

$$\begin{aligned} &\mathbb{P}[\ell(\mathsf{f}_n^*(X^n)) \ge K_n]\\ &\quad\le \mathbb{P}[\imath_{X^n}(X^n) \ge K_n] && (221)\\ &\quad\le Q\Big(Q^{-1}(\epsilon) + \frac{C}{\sigma\sqrt{n}}\Big) + \frac{A}{\sqrt{n}}, && (222)\end{aligned}$$

where (222) follows from Theorem 21. Since

$$Q'(x) = -\phi(x) \tag{223}$$

$$0 \le Q''(x) = x\phi(x) \le \frac{1}{\sqrt{2\pi e}}, \quad x \ge 0, \tag{224}$$

a second-order Taylor expansion of the first term in the right side of (222) gives

$$\begin{aligned} &\mathbb{P}[\ell(\mathsf{f}_n^*(X^n)) \ge K_n]\\ &\quad\le \epsilon - \frac{C}{\sigma\sqrt{n}}\left\{\phi(Q^{-1}(\epsilon)) - \frac{C}{2\sigma\sqrt{2\pi e n}} - \frac{A\sigma}{C}\right\}, && (225)\end{aligned}$$

and choosing C as in (194) for n satisfying (195) the right side of (225) is bounded above by ϵ. Therefore, $\mathbb{P}[\ell(\mathsf{f}_n^*(X^n)) > K_n] \le \epsilon$, which, by definition implies that $nR^*(n,\epsilon) \le K_n$, as claimed.

Proof of Theorem 20: Applying Theorem 4 with X^n in place of X and with $\delta > 0$ and $K_n \ge 1$ arbitrary, we obtain

$$\begin{aligned} &\mathbb{P}[\ell(\mathsf{f}_n^*(X^n)) \ge K_n]\\ &\quad\ge \mathbb{P}\Big[\imath_{X^n}(X^n) \ge K_n + \delta\Big] - 2^{-\delta} && (226)\\ &\quad\ge Q\Big(\frac{K_n - nH(\mathbf{X}) + \delta}{\sigma\sqrt{n}}\Big) - \frac{A}{\sqrt{n}} - 2^{-\delta}, && (227)\end{aligned}$$

where (227) now follows from Theorem 21. Letting $\delta = \delta_n = \frac{1}{2}\log_2 n$ and

$$K_n = nH(\mathbf{X}) + \sigma\sqrt{n}Q^{-1}(\epsilon) - \delta - \frac{\sigma(A+1)}{\phi(Q^{-1}(\epsilon))}, \tag{228}$$

yields

$$\begin{aligned} &\mathbb{P}[\ell(\mathsf{f}_n^*(X^n)) \ge K_n]\\ &\quad\ge Q\Big(Q^{-1}(\epsilon) - \frac{(A+1)}{\phi(Q^{-1}(\epsilon))\sqrt{n}}\Big) - \frac{A+1}{\sqrt{n}}. && (229)\\ &\quad> \epsilon && (230)\end{aligned}$$

which holds as long as $\epsilon \in (0, 1/2)$ and

$$n \ge \left(\frac{A+1}{Q^{-1}(\epsilon)\phi(Q^{-1}(\epsilon))}\right)^2, \tag{231}$$

in order to ensure that the argument of the Q function in (229) is nonnegative. Note that in (229) we have used a two-term Taylor expansion of Q based on $Q^{-1}(\epsilon) > 0$ and $Q'(x) = -\phi(x)$.

We conclude from (230) that $nR^*(n,\epsilon) > K_n - 1$, as claimed.

VII. Source Dispersion and Varentropy

Traditionally, refined analyses in lossless data compression have focused attention on the *redundancy*, defined as the difference between the minimum average compression rate and the entropy rate. As mentioned in Section I-E, the per-symbol redundancy is positive and of order $O\left(\frac{1}{n}\right)$ when the prefix condition is enforced, while it is $-\frac{1}{2n}\log_2 n + O(\frac{1}{n})$ without the prefix condition. But since the results in Sections V and VI show that the standard deviation of the best achievable compression rate is of order $O(\frac{1}{\sqrt{n}})$, the deviation of the rate from the entropy will be dominated by these stochastic fluctuations. Therefore, as noted in [19], it is of primary importance to analyze the variance of the optimal codelengths. To that end, we introduce the following operational definition:

Definition 2: The dispersion D (measured in bits^2) of a source $\mathbf{X}$ is

$$D = \limsup_{n\to\infty} \frac{1}{n}\mathsf{Var}(\ell(\mathsf{f}_n^*(X^n))), \tag{232}$$

where $\ell(\mathsf{f}_n^*(\cdot))$ is the length of the optimum fixed-to-variable lossless code (cf. Section I-B).

As we show in Theorem 23 below, for a broad class of sources, the dispersion D is equal to the source varentropy rate $\sigma^2(\mathbf{X})$ defined in (121). Moreover, in view of the Gaussian approximation bounds for $R^*(n,\epsilon)$ in Sections V and VI – 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 $n^*(R,\epsilon)$ in Section I-B, 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

$$n^*((1+\eta)H(\mathbf{X}),\epsilon) \approx \frac{\sigma^2(\mathbf{X})}{H^2(\mathbf{X})}\left(\frac{Q^{-1}(\epsilon)}{\eta}\right)^2 \tag{233}$$

$$= \frac{D}{H^2(\mathbf{X})}\left(\frac{Q^{-1}(\epsilon)}{\eta}\right)^2, \tag{234}$$

i.e., by the product of a factor that depends only on the source (through $\sigma^2(\mathbf{X})/H^2(\mathbf{X})$), and a factor that depends only on the design requirements ϵ and η. Note the close parallel with the notion of channel dispersion introduced in [32].

Example 5: Coin flips with bias p have varentropy,

$$\sigma^2(\mathbf{X}) = p(1-p)\log^2\frac{1-p}{p}, \tag{235}$$

so the key parameter in (234) which characterizes the time horizon required for the source to become "typical" is

$$\frac{\sigma^2(\mathbf{X})}{H^2(\mathbf{X})} = \frac{D}{H^2(\mathbf{X})} = p(1-p)\left(\frac{\log_2\frac{1-p}{p}}{h(p)}\right)^2 \tag{236}$$

where $h(\cdot)$ denotes the binary entropy function in bits. In view of Example 4, the normalized dispersion for the binary symmetric Markov chain with transition probability p is also given by (236).

Example 6: For a memoryless source whose marginal is the geometric distribution,

$$P_X(k) = q(1-q)^k, \quad k \geq 0, \tag{237}$$

the ratio of varentropy to squared entropy is

$$\frac{\sigma^2(\mathbf{X})}{H^2(\mathbf{X})} = \frac{D}{H^2(\mathbf{X})} = (1-q)\left(\frac{\log_2(1-q)}{h(q)}\right)^2. \tag{238}$$

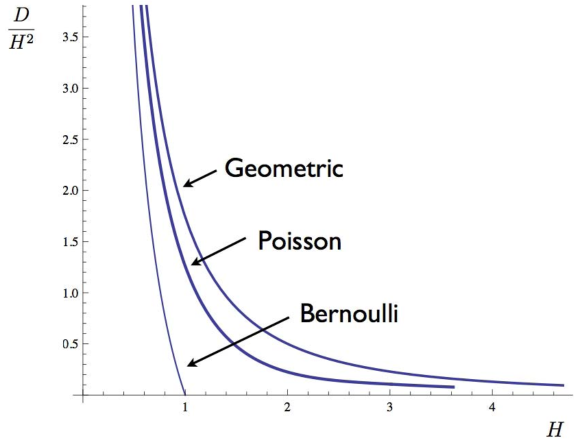

Fig. 4. Normalized dispersion as a function of entropy for memoryless sources.

Figure 4 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 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

$$\imath_{X^n}(x^n) \leq An. \tag{239}$$

Any memoryless source belongs to the class of linear information growth. Also note that every irreducible and aperiodic Markov chain is a linear information growth source: Writing q for the smallest nonzero element of the transition matrix, and π for the smallest nonzero probability for X_1, we easily see that (239) is satisfied with $N_0 = 2$, $A = \log_2(1/q) + |\log_2(q/\pi)|$. The class of linear information growth sources is related, at least at the level of intuition, to the class of finite-energy processes considered by Shields [35] and to processes satisfying the Doeblin-like condition of Kontoyiannis and Suhov [21].

We proceed to show an interesting regularity result for linear information growth sources:

Lemma 1: Suppose $\mathbf{X}$ is a (not necessarily stationary or ergodic) linear information growth source. Then:

$$\lim_{n\to\infty}\frac{1}{n}\mathbb{E}\Big[\Big(\ell(\mathsf{f}_n^*(X^n)) - \imath_{X^n}(X^n)\Big)^2\Big] = 0. \tag{240}$$

Proof: For brevity, denote $\ell_n = \ell(\mathsf{f}_n^*(X^n))$ and $\imath_n = \imath_{X^n}(X^n)$, respectively. Select an arbitrary τ_n. The expectation of interest is

$$\mathbb{E}[(\ell_n - \imath_n)^2] = \mathbb{E}[(\ell_n - \imath_n)^2 1\{\ell_n \geq \imath_n - \tau_n\}] + \mathbb{E}[(\ell_n - \imath_n)^2 1\{\ell_n < \imath_n - \tau_n\}]. \tag{241}$$

Since $\ell_n \leq \imath_n$ always, on the event $\{\ell_n \geq \imath_n - \tau_n\}$ we have $(\ell_n - \imath_n)^2 \leq \tau_n^2$. Also, by the linear information growth assumption we have the bound $0 \leq \imath_n - \ell_n \leq \imath_n \leq Cn$ for a finite constant C and all n large enough. Combining these two observations with Theorem 5, we obtain that

$$\mathbb{E}[(\ell_n - \imath_n)^2] \leq \tau_n^2 + C^2 n^2 \mathbb{P}\{\ell_n < \imath_n - \tau_n\} \tag{242}$$

$$\leq \tau_n^2 + C^2 n^2 2^{-\tau_n}\left(n \log_2 |\mathcal{A}| + 1\right) \tag{243}$$

$$\leq \tau_n^2 + C' n^3 2^{-\tau_n}, \tag{244}$$

for some $C' < \infty$ and all n large enough. Taking $\tau_n = 3\log_2 n$, dividing by n and letting $n \to \infty$ gives the claimed result. ■

Note that we have actually proved a stronger result, namely,

$$\mathbb{E}\left[\left(\ell(\mathsf{f}_n^*(X^n)) - \imath_{X^n}(X^n)\right)^2\right] = O(\log^2 n). \tag{245}$$

Linear information growth is sufficient for dispersion to equal varentropy rate:

Theorem 22: If the source $\mathbf{X}$ has linear information growth and finite varentropy rate, then:

$$D = \sigma^2(\mathbf{X}). \tag{246}$$

Proof: For notational convenience, we abbreviate $H(X^n)$ as H_n, and as in the previous proof we write, $\ell_n = \ell(\mathsf{f}_n^*(X^n))$ and $\imath_n = \imath_{X^n}(X^n)$. Expanding the definition of the variance of ℓ_n, we obtain,

$$\begin{aligned}&\mathsf{Var}(\ell(\mathsf{f}_n^*(X^n)))\\ &= \mathbb{E}\left[(\ell_n - \mathbb{E}[\ell_n])^2\right] &(247)\\ &= \mathbb{E}\left[\left((\ell_n - \imath_n) + (\imath_n - H_n) - \mathbb{E}[\ell_n - \imath_n]\right)^2\right] &(248)\\ &= \mathbb{E}[(\ell_n - \imath_n)^2] + \mathbb{E}[(\imath_n - H_n)^2] - \mathbb{E}^2[\ell_n - \imath_n]\\ &\quad +2\,\mathbb{E}[(\ell_n - \imath_n)(\imath_n - H_n)], &(249)\end{aligned}$$

and therefore, using the Cauchy-Schwarz inequality twice,

$$\begin{aligned}&|\mathsf{Var}(\ell(\mathsf{f}_n^*(X^n))) - \mathsf{Var}(\imath_{X^n}(X^n))|\\ &= |\mathbb{E}[(\ell_n - \imath_n)^2] - \mathbb{E}^2[\ell_n - \imath_n]\\ &\quad +2\,\mathbb{E}[(\ell_n - \imath_n)(\imath_n - H_n)]| &(250)\\ &\leq 2\,\mathbb{E}[(\ell_n - \imath_n)^2]\\ &\quad +2\sqrt{\mathbb{E}[(\ell_n - \imath_n)^2]}\sigma(\imath_{X^n}(X^n)). &(251)\end{aligned}$$

where $\sigma(\cdot)$ indicates the standard deviation. Dividing by n and letting $n \to \infty$, we obtain that the first term tends to zero by Lemma 1, and the second term becomes the square root of

$$\frac{4}{n}\mathbb{E}[(\ell_n - \imath_n)^2] \times \frac{1}{n}\mathsf{Var}(\imath_{X^n}(X^n)) \tag{252}$$

which also tends to zero by Lemma 1 and the finite-varentropy rate assumption. Therefore,

$$\lim_{n\to\infty} \frac{1}{n}|\mathsf{Var}(\ell(\mathsf{f}_n^*(X^n))) - \mathsf{Var}(\imath_{X^n}(X^n))| = 0, \tag{253}$$

which, in particular, implies that $\sigma^2(\mathbf{X}) = D$. ■

In view of (245), if we normalize by $\sqrt{n}\log n$, instead of n in the last step of the proof of Theorem 22, we obtain the stronger result:

$$|\mathsf{Var}(\ell(\mathsf{f}_n^*(X^n))) - \mathsf{Var}(\imath_{X^n}(X^n)))| = O\left(\sqrt{n}\log_2 n\right). \tag{254}$$

Also, Lemma 1 and Theorem 22 remain valid if instead of the linear information growth condition we invoke the weaker assumption that there exists a sequence $\epsilon_n = o(\sqrt{n})$, such that

$$\max_{x^n:\ P_{X^n}(x^n)\neq 0} \imath_{X^n}(x^n) = o\left(2^{\epsilon_n/2}\right). \tag{255}$$

For Markov chains, the various results satisfied by varentropy and dispersion are as follows.

Theorem 23: Let $\mathbf{X}$ be an irreducible, aperiodic (not necessarily stationary) Markov source with entropy rate $H(\mathbf{X})$. Then:

1) The varentropy rate $\sigma^2(\mathbf{X})$ defined in (121) exists as the limit

$$\sigma^2(\mathbf{X}) = \lim_{n\to\infty} \frac{1}{n}\mathsf{Var}(\imath_{X^n}(X^n))). \tag{256}$$

2) The dispersion D defined in (232) exists as the limit

$$D = \lim_{n\to\infty} \frac{1}{n}\mathsf{Var}(\ell(\mathsf{f}_n^*(X^n))). \tag{257}$$

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

$$\sigma^2(\mathbf{X}) = \lim_{\epsilon\to 0}\lim_{n\to\infty} \frac{n\,(R^*(n,\epsilon) - H(\mathbf{X}))^2}{2\ln\frac{1}{\epsilon}} \tag{258}$$

$$= \lim_{\epsilon\to 0}\lim_{n\to\infty} n\left(\frac{R^*(n,\epsilon) - H(\mathbf{X})}{Q^{-1}(\epsilon)}\right)^2, \tag{259}$$

as long as $\sigma^2(\mathbf{X})$ is nonzero.

Proof: The limiting expression in part 1) was already established in Theorem 13 of Section IV; see also the discussion leading to (191) in Section VI. Recalling that every irreducible and aperiodic Markov source is a linear information growth source, combining part 1) with Theorem 22 immediately yields the results of parts 2) and 3).

Finally, part 4) follows from the results of Section VI. Under the present assumptions, Theorems 19 and 20 together imply that there is a finite constant C_1 such that

$$\begin{aligned}&\left|\sqrt{n}(R^*(n,\epsilon) - H(\mathbf{X})) - \sigma(\mathbf{X})Q^{-1}(\epsilon)\right|\\ &\leq \frac{1}{2}\frac{\log_2 n}{\sqrt{n}} + \frac{C_1}{\sqrt{n}}, &(260)\end{aligned}$$

for all $\epsilon \in (0, 1/2)$ and all n large enough. Therefore,

$$\lim_{n\to\infty} n(R^*(n,\epsilon) - H(\mathbf{X}))^2 = \sigma^2(\mathbf{X})(Q^{-1}(\epsilon))^2. \tag{261}$$

Dividing by $2\ln\frac{1}{\epsilon}$, letting $\epsilon \downarrow 0$, and recalling the simple fact that $(Q^{-1}(\epsilon))^2 \sim 2\ln\frac{1}{\epsilon}$ (see, e.g., [39, Section 3.3]) proves (259) and completes the proof of the theorem. ■

From Theorem 22 it follows that, for a broad class of sources including all ergodic Markov chains with nonzero varentropy rate,

$$\lim_{n\to\infty} \frac{\mathsf{Var}\left(\ell(\mathsf{f}_n^*(X^n))\right)}{\mathsf{Var}\left(\imath_{X^n}(X^n)\right)} = 1. \tag{262}$$

Analogously to Theorem 10, we could explore whether (262) might hold under broader conditions, including the general setting of possibly non-serial sources. However, consider the following simple example.

Example 7: As in Example 3, let X_M be equiprobable on a set of M elements, then,

$$H(X_M) = \log_2 M \tag{263}$$

$$\mathsf{Var}\left(\imath_{X_M}(X_M)\right) = 0 \tag{264}$$

$$\limsup_{M\to\infty} \mathsf{Var}\left(\ell(\mathsf{f}^*(X_M))\right) = 2 + \frac{1}{4} \tag{265}$$

$$\liminf_{M\to\infty} \mathsf{Var}\left(\ell(\mathsf{f}^*(X_M))\right) = 2. \tag{266}$$

To verify (265) and (266), define the function,

$$s(K) = \sum_{i=1}^{K} i^2\, 2^i \tag{267}$$

$$= -6 + 2^{K+1}(3 - 2K + K^2). \tag{268}$$

It is straightforward to check that

$$\mathbb{E}\left[\ell^2(\mathsf{f}^*(X_M))\right] = \frac{1}{M}\Big(s(\lfloor\log_2 M\rfloor) - (\lfloor\log_2 M\rfloor)^2 \times \left(2^{\lfloor\log_2 M\rfloor+1} - M - 1\right) \tag{269}$$

Together with (102), (269) results in,

$$\mathsf{Var}\left(\ell(\mathsf{f}^*(X_M))\right) = 3\xi_M - \xi_M^2 + o(1), \tag{270}$$

with

$$\xi_M = \frac{2^{1+\lfloor\log_2 M\rfloor}}{M}, \tag{271}$$

which takes values in $(1, 2]$. On that interval, the parabola $3x - x^2$ takes a minimum value of 2 and a maximum value of $(3/2)^2$, and (265), (266) follow.

Although the ratio of optimal codelength variance to the varentropy rate may be infinity as illustrated in Example 7, we do have the following counterpart of the first-moment result in Theorem 10 for the second moments:

Theorem 24: For any (not necessarily serial) source $\mathbf{X} = \{P_{X^{(n)}}\}$,

$$\lim_{n\to\infty} \frac{\mathbb{E}[\ell^2(\mathsf{f}_n^*(X^{(n)}))]}{\mathbb{E}\left[\imath^2_{X^{(n)}}(X^{(n)})\right]} = 1, \tag{272}$$

as long as the denominator diverges.

Proof: Theorem 3 implies that

$$\mathbb{E}[\ell^2(\mathsf{f}_n^*(X^{(n)}))] \leq \mathbb{E}\left[\imath^2_{X^{(n)}}(X^{(n)})\right]. \tag{273}$$

Therefore, the lim sup in (272) is bounded above by 1. To establish the corresponding lower bound, fix an arbitrary $\vartheta > 0$. Then,

$$\begin{aligned} &\mathbb{E}[\ell^2(\mathsf{f}_n^*(X^{(n)}))] \\ &= \sum_{k\geq 1} \mathbb{P}\left[\ell^2(\mathsf{f}_n^*(X^{(n)})) \geq k\right] && (274)\\ &= \sum_{k\geq 1} \mathbb{P}\left[\ell_n^* \geq \sqrt{k}\right] && (275)\\ &= \sum_{k\geq 1} \mathbb{P}\left[\ell_n^* \geq \lceil\sqrt{k}\rceil\right] && (276)\\ &\geq \sum_{k\geq 1} \left[\mathbb{P}\left[\imath_{X^{(n)}}(X^{(n)}) \geq (1+\vartheta)\lceil\sqrt{k}\rceil\right] - 2^{-\vartheta\lceil\sqrt{k}\rceil}\right], && (277) \end{aligned}$$

where (277) follows by letting $\tau = \vartheta\lceil\sqrt{k}\rceil$ in the converse Theorem 4. Therefore,

$$\begin{aligned} &\mathbb{E}[\ell^2(\mathsf{f}_n^*(X^{(n)}))] \\ &\geq -C_\vartheta + \sum_{k\geq 1} \mathbb{P}\left[\imath^2_{X^{(n)}}(X^{(n)}) \geq (1+\vartheta)^2\lceil\sqrt{k}\rceil^2\right] && (278)\\ &\geq -D_\vartheta + \sum_{k\geq 1} \mathbb{P}\left[\frac{\imath^2_{X^{(n)}}(X^{(n)})}{(1+\vartheta)^3} \geq k\right] && (279)\\ &\geq -D_\vartheta + \frac{1}{(1+\vartheta)^3}\mathbb{E}[\imath^2_{X^{(n)}}(X^{(n)})]. && (280) \end{aligned}$$

where C_ϑ, D_ϑ are positive scalars that only vary with ϑ. Note that (278) holds because $a^{\sqrt{k}}$ is summable for all $0 < a < 1$; (279) holds because $(1+\vartheta)k \geq \lceil\sqrt{k}\rceil^2$ for all sufficiently large k; and (280) holds because the mean of a nonnegative random variable is the integral of the complementary cumulative distribution function, which in turn satisfies

$$\int_k^{k+1} (1 - F(x))\, dx \geq 1 - F(k+1), \tag{281}$$

Dividing both sides of (278)-(280) by the second moment $\mathbb{E}[\imath^2_{X^{(n)}}(X^{(n)})]$ and letting $n \to \infty$, we conclude that the ratio in (272) is bounded below by $(1+\vartheta)^{-3}$. Since ϑ can be taken to be arbitrarily small, this proves that the lim inf (272) is bounded below by 1, as required. ■

References

[1] N. Alon and A. Orlitsky, "A lower bound on the expected length of one-to-one codes," *IEEE Trans. Inf. Theory*, vol. 40, no. 5, pp. 1670–1672, Sep. 1994.

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

[3] C. Blundo and R. de Prisco, "New bounds on the expected length of one-to-one codes," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 246–250, Jan. 1996.

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

[5] J. Cheng, T. K. Huang, and C. Weidmann, "New bounds on the expected length of optimal one-to-one codes," *IEEE Trans. Inf. Theory*, vol. 53, no. 5, pp. 1884–1895, May 2007.

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

[7] K. L. Chung, *A Course in Probability Theory*, 2nd ed. New York, NY, USA: Academic, 1974.

[8] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006

[9] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[10] A. Dembo and I. Kontoyiannis, "Source coding, large deviations, and approximate pattern matching," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1590–1615, Jun. 2002.

[11] J. G. Dunham, "Optimal noiseless coding of random variables (Corresp.)," *IEEE Trans. Inf. Theory*, vol. 26, no. 3, p. 345, May 1980.

[12] P. Hall, *Rates of Convergence in the Central Limit Theorem*. London, U.K.: Pitman, 1982.

[13] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 752–772, May 1993.

[14] T. C. Hu, D. J. Kleitman, and J. K. Tamaki, "Binary trees optimum under various criteria," *SIAM J. Appl. Math.*, vol. 37, no. 2, pp. 246–256, 1979.

[15] P. Humblet, "Generalization of Huffman coding to minimize the probability of buffer overflow (Corresp.)," *IEEE Trans. Inf. Theory*, vol. 27, no. 2, pp. 230–232, Mar. 1981.

[16] I. A. Ibragimov, "Some limit theorems for stationary processes," *Theory Probab. Appl.*, vol. 7, no. 4, pp. 349–382, 1962.

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

[18] J. C. Kieffer, "Sample converses in source coding theory," *IEEE Trans. Inf. Theory*, vol. 37, no. 2, pp. 263–268, Mar. 1991.

[19] I. Kontoyiannis, "Second-order noiseless source coding theorems," *IEEE Trans. Inf. Theory*, vol. 43, no. 3, pp. 1339–1341, Jul. 1997.

[20] I. Kontoyiannis, "Asymptotic recurrence and waiting times for stationary processes," *J. Theoretical Probab.*, vol. 11, no. 3, pp. 7950–811, 1998.

[21] I. Kontoyiannis and Y. M. Suhov, "Prefixes and the entropy rate for long-range sources," *Probability Statistics and Optimization*, F. P. Kelly, Ed. New York, NY, USA: Wiley, 1994.

[22] V. Yu. Korolev and I. G. Shevtsova, "On the upper bound for the absolute constant in the Berry–Eséen inequality," *Theory Probab. Appl.*, vol. 54, no. 4, pp. 638–658, 2010.

[23] S. K. Leung-Yan-Cheong and T. Cover, "Some equivalences between Shannon entropy and Kolmogorov complexity," *IEEE Trans. Inf. Theory*, vol. 24, no. 3, pp. 331–338, May 1978.

[24] B. Mann, "Berry–Eséen central limit theorems for Markov chains," Ph.D. dissertation, Dept. Math., Harvard Univ., Cambridge, MA, USA, 1996.

[25] B. McMillan, "The basic theorems of information theory," *Ann. Math. Statist.*, vol. 24, pp. 196–219, Jun. 1953.

[26] B. McMillan, "Two inequalities implied by unique decipherability," *IRE Trans. Inf. Theory*, vol. 2, no. 4, pp. 115–116, Dec. 1956.

[27] N. Merhav, "Universal coding with minimum probability of codeword length overflow," *IEEE Trans. Inf. Theory*, vol. 37, no. 3, pp. 556–563, May 1991.

[28] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[29] S. V. Nagaev, "More exact limit theorems for homogeneous Markov chains," *Theory Probab. Appl.*, vol. 6, no. 1, pp. 62–81, 1961.

[30] V. V. Petrov, *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*. Oxford, U.K.: Oxford Science, 1995.

[31] W. Philipp and W. Stout, *Almost Sure Invariance Principles for Partial Sums of Weakly Dependent Random Variables*. Providence, RI, USA: AMS, 1975.

[32] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[33] J. Rissanen, "Tight lower bounds for optimum code length," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 348–349, Mar. 1982.

[34] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 4, pp. 623–656, 1948.

[35] P. C. Shields, *The Ergodic Theory of Discrete Sample Paths* (Graduate Studies in Mathematics). Providence, RI, USA: AMS, 1996.

[36] G. Somasundaram and A. Shrivastava, *Information Storage and Management: Storing, Managing, and Protecting Digital Information in Classic, Virtualized, and Cloud Environments*. New York, NY, USA: Wiley, 2012.

[37] V. Strassen, "Asymptotische abschäzungen in Shannons informationstheorie," in *Proc. Trans. Third Prague Conf. Inf. Theory, Statist., Decision Funct., Random Process.*, 1964, pp. 689–723.

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

[39] S. Verdú, *Multiuser Detection*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[40] S. Verdú, "teaching it," presented at the IEEE Int. Symp. Inf. Theory, Nice, France, Jun. 2007.

[41] S. Verdú, "Teaching lossless data compression," *IEEE Inf. Theory Soc. Newsletter*, vol. 61, no. 1, pp. 18–19, Apr. 2011.

[42] E. I. Verriest, "An achievable bound for optimal noiseless coding of a random variable," *IEEE Trans. Inf. Theory*, vol. 32, no. 4, pp. 592–594, Jul. 1986.

[43] A. D. Wyner, "An upper bound on the entropy series," *Inf. Control*, vol. 20, no. 2, pp. 176–181, 1972.

[44] A. A. Yushkevich, "On limit theorems connected with the concept of entropy of Markov chains," *Uspekhi Mat. Nauk*, vol. 8, no. 5(57), pp. 177–180, 1953.

Ioannis Kontoyiannis was born in Athens, Greece, in 1972. He received the B.Sc. degree in mathematics in 1992 from Imperial College (University of London), U.K., and in 1993 he obtained a distinction in Part III of the Cambridge University Pure Mathematics Tripos. He received the M.S. degree in statistics and the Ph.D. degree in electrical engineering, both from Stanford University, Stanford, CA, in 1997 and 1998, respectively.

From June 1998 to August 2001, he was an Assistant Professor with the Department of Statistics, Purdue University, West Lafayette, IN (and also, by courtesy, with the Department of Mathematics, and the School of Electrical and Computer Engineering). From August 2000 until July 2005, he was an Assistant, then Associate Professor, with the Division of Applied Mathematics and with the Department of Computer Science, Brown University, Providence, RI. Since March 2005, he has been with the Department of Informatics, Athens University of Economics and Business, where he is currently a Professor.

Dr. Kontoyiannis was awarded the Manning endowed Assistant Professorship in 2002, and was awarded an honorary Master of Arts degree Ad Eundem, in 2005, both by Brown University. In 2004, he was awarded a Sloan Foundation Research Fellowship. He has served two terms as an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY. His research interests include data compression, applied probability, information theory, statistics, simulation, and mathematical biology.

Sergio Verdú (S'80–M'84–SM'88–F'93) is on the faculty of the School of Engineering and Applied Science at Princeton University. A member of the National Academy of Engineering, Verdú is the recipient of the 2007 Claude E. Shannon Award and of the 2008 IEEE Richard W. Hamming Medal.