Lempel-Ziv Complexity, Empirical Entropies, and Chain Rules

Neri Merhav

The Andrew & Erna Viterbi Faculty of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, ISRAEL
E–mail: merhav@ee.technion.ac.il

Abstract

We derive upper and lower bounds on the overall compression ratio of the 1978 Lempel-Ziv (LZ78) algorithm, applied independently to k-blocks of a finite individual sequence. Both bounds are given in terms of normalized empirical entropies of the given sequence. For the bounds to be tight and meaningful, the order of the empirical entropy should be small relative to k in the upper bound, but large relative to k in the lower bound. Several non-trivial conclusions arise from these bounds. One of them is a certain form of a chain rule of the Lempel-Ziv (LZ) complexity, which decomposes the joint LZ complexity of two sequences, say, $\boldsymbol{x}$ and $\boldsymbol{y}$, into the sum of the LZ complexity of $\boldsymbol{x}$ and the conditional LZ complexity of $\boldsymbol{y}$ given $\boldsymbol{x}$ (up to small terms). The price of this decomposition, however, is in changing the length of the block. Additional conclusions are discussed as well.

1 Introduction

In the second half of the 1970s, Jacob Ziv and Abraham Lempel introduced a transformative concept in information theory [8], [9], [12]. Departing from traditional probabilistic frameworks which typically assumed memoryless sources and channels with well-defined statistical characteristics, they proposed a novel perspective known as the *individual-sequence approach*. This approach, when coupled with models of finite-state (FS) encoders and decoders, opened up a new avenue for understanding universal data compression and coded communication. Within this innovative framework, the foundational ideas of what would become the Lempel-Ziv (LZ) algorithms began to take shape, culminating in the development of the LZ77 and LZ78 algorithms in 1977 and 1978, respectively, as well as quite a few other variants. These algorithms have since become iconic in the field, celebrated not only for their theoretical elegance, but also for their exceptional

practical utility. The impact of the LZ family of algorithms, including its subsequent variations, has been far-reaching, deeply embedded in the everyday technologies that rely on digital storage and communication, from computers and smartphones to virtually every device that handles digital data.

In subsequent years, the individual-sequence framework was extended along various lines of study. One notable development appeared in [10], where Ziv examined a fixed-rate coding scenario involving side information, with both the source and the side information sequences being deterministic (individual) sequences. Along this setting, with both the encoder and decoder being modeled as finite-state machines, he introduced and rigorously characterized the concept of fixed-rate conditional complexity. This measure captures the minimum achievable rate for almost-lossless compression of a source sequence given a side information sequence. Remarkably, echoing the classical result from Slepian-Wolf coding [6], Ziv demonstrated that access to side information at the encoder is not necessary in order to attain this conditional complexity. The following year, in [11], Ziv proposed a variable-rate counterpart to the conditional Lempel-Ziv (LZ) complexity in a markedly different context, serving as a universal decoding metric for unknown finite-state channels. This conditional complexity measure later garnered attention for its applicability to source coding with side information, as explored further in [3], [7], and more recently in [4].

Just as LZ complexity serves as the individual-sequence analogue of the entropy rate in the probabilistic setting, the conditional LZ complexity naturally parallels the conditional entropy rate. Following this line of analogy between the probabilistic and individual-sequence frameworks, a compelling question arises: Does the LZ complexity measure obey a chain rule? That is, can the joint LZ complexity of a sequence pair, say, $(\boldsymbol{x}, \boldsymbol{y})$, be decomposed into the sum of the LZ complexity of $\boldsymbol{x}$ and the conditional LZ complexity of $\boldsymbol{y}$ given $\boldsymbol{x}$, or, symmetrically, the reverse?

On the face of it, a close examination of the mathematical expressions of these three complexity measures for finite-length sequences offers very little reason to hope for an affirmative answer to this question. Surprisingly, however, such a chain-rule decomposition was shown to hold at least in a specific sense of the asymptotic regime of infinitely long sequences [4]. Given the central role that the chain rule for Shannon entropy plays in classical information theory, it is natural to envision that an analogous chain rule for LZ complexity could emerge as a foundational principle in the

development of an information theory tailored to individual sequences.

Consider, for instance, the problem of separately compressing almost losslessly and jointly decompressing two individual source sequences, in the spirit of Slepian-Wolf coding [6], but with the limitation that only finite-state encoders are allowed. As explored in [4], characterizing the achievable rate region in this setting brings forth a fundamental question. In the classical probabilistic framework, for two correlated discrete memoryless sources X and Y, the achievable rate region is well understood. It is defined by the set of rate pairs, $\{(R_{\mathrm{x}}, R_{\mathrm{y}}) : \ R_{\mathrm{x}} \geq H(X|Y), \ R_{\mathrm{y}} \geq H(Y|X), \ R_{\mathrm{x}} + R_{\mathrm{y}} \geq H(X,Y)\}$, where the corner points, $(H(X), H(Y|X))$ and $(H(X|Y), H(Y))$, arise naturally due to the chain rule of entropy, i.e., $H(X) = H(X,Y) - H(Y|X)$ and $H(Y) = H(X,Y) - H(X|Y)$. In the individual-sequence framework, as described in [4], a similar region can be defined, this time, replacing entropic quantities with their corresponding LZ complexity counterparts: the conditional complexities of the sequences and their joint complexity. However, unlike the probabilistic case, it is not immediately clear a-priori whether a chain rule exists that allows for a decomposition of the joint LZ complexity into unconditional components, such as the individual complexities of $\boldsymbol{x}$ and $\boldsymbol{y}$ in a manner analogous to the marginal entropy components of the corner points. In this context, the existence of a chain rule for LZ complexities would be, not only natural, but also instrumental in shaping a deeper understanding of compression limits for individual sequences.

Another illustrative example involves the concept of successive refinement for individual sequences, as explored in [5]. In the system model considered there, the encoder architecture comprises two main components: a reproduction encoder and a cascaded finite-state lossless encoder. The reproduction encoder generates two distorted versions of the source, one providing a coarse approximation with relatively high distortion, and the other offering a finer, more accurate representation with reduced distortion. These two reproduction vectors are then passed to the lossless encoder, which produces two compressed bit-streams that, taken together, represent both reproductions without introducing any additional distortion. The first bit-stream corresponds to the coarse description, and ideally, it alone captures the LZ complexity of the coarse reproduction. Together, the two bit-streams are expected to match the joint LZ complexity of both reproduction vectors. Since the first stream compresses only the coarse version, it is most natural for the second-stage encoder to compress the refined reproduction using the coarse one as side information. Consequently,

achieving the overall joint LZ complexity hinges on the existence of a chain rule for LZ complexity, at least in an asymptotic sense.

Earlier, we mentioned that in [4] a certain form of an asymptotic chain rule of LZ complexities was established for infinite individual sequences (more details will be provided in Section 4). Our main result in this work is another form of a chain rule that applies even to finite sequences, and it is therefore, stronger, more refined, and more explicit. To this end, we first derive upper and lower bounds on the overall compression ratio of the LZ algorithm, applied independently to k-blocks of a finite individual sequence. Both bounds are given in terms of normalized empirical entropies of the given sequence. For the bounds to be tight and meaningful, the order the empirical entropy should be small relative to k in the upper bound, but large relative to k in the lower bound. Several non-trivial conclusions arise from these bounds. One of them is the above mentioned chain rule of the Lempel-Ziv (LZ) complexity, which decomposes the joint LZ complexity of two sequences into the sum of the LZ complexity of one sequence and the conditional LZ complexity of the other sequence given the former (up to small terms). The price of this decomposition, however, is in changing the length of the block. Additional conclusions are discussed as well.

Finally, it is interesting to point out that the Kolmogorov complexity also obeys a certain approximate chain rule (up to a certain redundancy term that vanishes as the sequence length grows), as asserted in the Kolmogorov-Levin theorem [2], [13].

The outline of the remaining part of this work is as follows. In Section 2, we establish notation conventions and provide some background. In Section 3, we derive upper and lower bounds on the average LZ complexity of k-blocks of a given individual sequence, in terms of empirical entropies, and finally, in Section 4, we provide the chain rule results, which are upper and lower bounds on the average of the LZ complexities of sequence pairs, in terms of the average of the chain-rule decompositions in a sense that will be made clear in the sequel. We end this paper with a comparison to the above-mentioned earlier derived chain rule for the LZ complexity which appears in [4].

2 Notation Conventions and Background

2.1 Notation Conventions

Throughout this article, we adopt the following notational conventions. Scalar random variables (RVs) will be represented by uppercase letters, their realizations by the corresponding lowercase letters, and their alphabets by calligraphic letters. The same convention extends to random vectors and their realizations, which will be denoted using superscripts to indicate dimension. For instance, X^m, m being a positive integer, denotes the random vector $(X_1, \ldots, X_m)$ and $(x_1, \ldots, x_m)$ represents a specific realization in $\mathcal{X}^m$, the m-fold Cartesian power of the alphabet $\mathcal{X}$. Segment notation will follow accordingly: for positive integers i and j, $i \leq j$, x_i^j and X_i^j denote the substrings $(x_i, x_{i+1}, \ldots, x_j)$ and $(X_i, X_{i+1}, \ldots, X_j)$, respectively. When $i = 1$ the subscript '1' is omitted for brevity. If $i > j$, both x_i^j and X_i^j refer to the empty string. Unless stated otherwise, all logarithms and exponentials are taken to base 2. The indicator function of an event $\mathcal{E}$ is denoted by $I\{\mathcal{E}\}$, that is, $I\{\mathcal{E}\} = 1$ if $\mathcal{E}$ occurs and $I\{\mathcal{E}\} = 0$ if not.

In the sequel, $x^n = (x_1, \ldots, x_n)$ and $y^n = (y_1, \ldots, y_n)$ and will designate individual sequences. The components, $\{x_i\}$ of x^n, and $\{y_i\}$ of y^n, all take values in the corresponding finite alphabets, $\mathcal{X}$ and $\mathcal{Y}$, whose cardinalities will be denoted by α and β, respectively. The infinite sequences $(x_1, x_2, \ldots)$ and $(y_1, y_2, \ldots)$ and will be denoted by $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively.

2.2 Background

2.2.1 Finite-State Encoders

Following the framework introduced in [12], we consider a model for lossless compression based on a finite-state encoder. Such an encoder is characterized by a quintuple

$$E = (\mathcal{X}, \mathcal{U}, \mathcal{Z}, f, g),$$

where: $\mathcal{X}$ denotes a finite input alphabet of cardinality $\alpha = |\mathcal{X}|$; $\mathcal{U}$ is a finite set of variable-length binary strings, possibly including the empty string λ (of length zero); $\mathcal{Z}$ is a finite set of internal states; $f : \mathcal{Z} \times \mathcal{X} \to \mathcal{U}$ is the output function, and $g : \mathcal{Z} \times \mathcal{X} \to \mathcal{Z}$ is the next-state transition function. Given an infinite input sequence $\boldsymbol{x} = (x_1, x_2, \ldots)$ with $x_i \in \mathcal{X}$, $i = 1, 2, \ldots$, henceforth referred to as the source sequence, the encoder E generates a corresponding infinite output sequence

$\boldsymbol{u} = (u_1, u_2, \ldots)$ with $u_i \in \mathcal{U}$, henceforth termed the compressed bit-stream, while simultaneously evolving through a sequence of internal states $\boldsymbol{z} = (z_1, z_2, \ldots)$ with $z_i \in \mathcal{Z}$. The system dynamics are governed recursively by the equations:

$$u_i = f(z_i, x_i), \tag{1}$$

$$z_{i+1} = g(z_i, x_i), \tag{2}$$

for $i = 1, 2, \ldots$, with a fixed initial state $z_1 = z_\star \in \mathcal{Z}$. If at any step $u_i = \lambda$, no output is produced, and this corresponds to encoder idling, where only the internal state is updated in response to the input symbol.

An encoder with s distinct internal states, henceforth referred to as an s-state encoder, is one for which $|\mathcal{Z}| = s$. For convenience, we adopt a few notational conventions from [12]: Given a segment of input symbols x_i^j with $i \leq j$ and an initial state z_i, the notation $f(z_i, x_i^j)$ denotes the corresponding segment of outputs u_i^j generated by the encoder E. Likewise, $g(z_i, x_i^j)$ denotes the resulting state z_{j+1} after processing the input segment x_i^j starting from state z_i.

A finite-state encoder E is said to be information lossless (IL) if, for every initial state $z_i \in \mathcal{Z}$, every positive integer n, and any input segment x_i^{i+n}, the triplet $(z_i, f(z_i, x_i^{i+n}), g(z_i, x_i^{i+n}))$ uniquely determines the original input segment x_i^{i+n}. In other words, the combination of the starting state, the resulting output sequence, and the final state after encoding is sufficient to fully reconstruct the input. Given an encoder E and an input sequence x^n, the compression ratio achieved by E on x^n is defined as

$$\rho_E(x^n) \triangleq \frac{L(u^n)}{n} = \frac{1}{n}\sum_{i=1}^{n} l(u_i) = \frac{1}{n}\sum_{i=1}^{n} l[f(z_i, x_i)] \tag{3}$$

where $L(u^n)$ denotes the total length (in bits) of the encoded binary string u^n, and $l(u_i)$ represents the length of the binary string $u_i = f(z_i, x_i)$ at each step i.

The class of all IL encoders $\{E\}$ with no more than s states is denoted by $\mathcal{E}(s)$. We next define the s-state compressibility of x^n by

$$\rho_s(x^n) = \min_{E \in \mathcal{E}(s)} \rho_E(x^n), \tag{4}$$

the asymptotic s-state compressibility of $\boldsymbol{x}$ by

$$\rho_s(\boldsymbol{x}) = \limsup_{n \to \infty} \rho_s(x^n), \tag{5}$$

and finally, the *finite–state compressibility* of $\boldsymbol{x}$ by

$$\rho(\boldsymbol{x}) = \lim_{s\to\infty} \rho_s(\boldsymbol{x}). \tag{6}$$

2.2.2 Empirical Distributions and Induced Information Measures

We define three types of empirical distributions of d-vectors (d – positive integer).

1. Assuming that d divides n, the empirical distribution pertaining to *non-overlapping blocks* of length d is defined as

$$\hat{P}_{\text{nob}}(z, w^d) \triangleq \frac{d}{n} \sum_{i=0}^{n/d-1} \mathcal{I}\{z_{id+1} = z,\ x_{id+1}^{id+d} = w^d\}, \quad z \in \mathcal{Z},\ w^d = (w_1, \ldots, w_d) \in \mathcal{X}^d. \tag{7}$$

2. The empirical distribution associated with a *sliding window* of length d is defined as

$$\hat{P}_{\text{sw}}(z, w^d) \triangleq \frac{1}{n-d+1} \sum_{i=0}^{n-d} \mathcal{I}\{z_{i+1} = z,\ x_{i+1}^{i+d} = w^d\}, \quad z \in \mathcal{Z},\ w^d \in \mathcal{X}^d. \tag{8}$$

3. The empirical distribution associated with a *cyclic sliding window* of length d is defined as

$$\hat{P}_{\text{csw}}(z, w^d) \triangleq \frac{1}{n} \sum_{i=0}^{n-1} \mathcal{I}\{z_{i+1} = z,\ x_{i+1}^{[(i-1)\oplus d]+1} = w^d\}, \quad z \in \mathcal{Z},\ w^d \in \mathcal{X}^d, \tag{9}$$

   where $\oplus$ denotes modulo-n addition.

Information measures associated with these empirical distributions will be denoted according to the conventional notation rules of the information theory literature, but with 'hats', with subscripts that indicate the type of the empirical distribution, and with notation of dependence on the data sequence x^n from which the statistics were gathered (using square brackets). For example, $\hat{H}_{\text{nob}}(X^d)[x^n]$ will denote the empirical entropy of an auxiliary random vector X^d that is governed by the empirical distribution, $\hat{P}_{\text{nob}}(\cdot)$ extracted from x^n.[1] Likewise, $\hat{H}_{\text{sw}}(X^d|Z)[x^n]$ will denote the empirical conditional entropy of an auxiliary random vector X^d given a random state variable Z that are drawn by the empirical distribution, $\hat{P}_{\text{sw}}(\cdot,\cdot)$, $\hat{I}_{\text{csw}}(X^d;Z)[x^n]$ will denote the empirical mutual information between the auxiliary random variables X^d and Z that are jointly disributed according to the empirical distribution, $\hat{P}_{\text{csw}}(\cdot)$, and so on.

[1] Note that there is no need to denote the dependence on z^n too since z^n is dictated by x^n for a given next-state function, g.

For the infinite sequence $\boldsymbol{x} = (x_1, x_2, \ldots)$, we define

$$\hat{H}_{\rm csw}(X^d)[\boldsymbol{x}] = \limsup_{n\to\infty} \hat{H}_{\rm csw}(X^d)[x^n]. \tag{10}$$

Similarly as shown in [12], for every $\boldsymbol{x}$, the sequence $\{\hat{H}_{\rm csw}(X^d)[\boldsymbol{x}]\}_{d\geq 1}$ is sub-additive as

$$\begin{aligned}
\hat{H}_{\rm csw}(X^{d_1+d_2})[x^n] &= \hat{H}_{\rm csw}(X^{d_1})[x^n] + \hat{H}_{\rm csw}(X_{d_1+1}^{d_1+d_2}|X^{d_1})[x^n] \\
&\leq \hat{H}_{\rm csw}(X^{d_1})[x^n] + \hat{H}_{\rm csw}(X_{d_1+1}^{d_1+d_2})[x^n] \\
&= \hat{H}_{\rm csw}(X^{d_1})[x^n] + \hat{H}_{\rm csw}(X^{d_2})[x^n],
\end{aligned} \tag{11}$$

and so, taking the limit superior of the left-most- and the right-most side, we have

$$\begin{aligned}
\hat{H}_{\rm csw}(X^{d_1+d_2})[\boldsymbol{x}] &= \limsup_{n\to\infty} \hat{H}_{\rm csw}(X^{d_1+d_2})[x^n] \\
&\leq \limsup_{n\to\infty} \{\hat{H}_{\rm csw}(X^{d_1})[x^n] + \hat{H}_{\rm csw}(X^{d_2})[x^n]\} \\
&\leq \limsup_{n\to\infty} \hat{H}_{\rm csw}(X^{d_1})[x^n] + \limsup_{n\to\infty} \hat{H}_{\rm csw}(X^{d_2})[x^n] \\
&= \hat{H}_{\rm csw}(X^{d_1})[\boldsymbol{x}] + \hat{H}_{\rm csw}(X^{d_2})[\boldsymbol{x}].
\end{aligned} \tag{12}$$

Consequently, the sequence $\{\frac{\hat{H}_{\rm csw}(X^d)[\boldsymbol{x}]}{d}\}_{d\geq 1}$ is convergent, and we shall denote

$$\bar{H}_{\rm csw}[\boldsymbol{x}] = \lim_{d\to\infty} \frac{\hat{H}_{\rm csw}(X^d)[\boldsymbol{x}]}{d}. \tag{13}$$

Returning to finite n, whenever the underlying sequence x^n is clear from the context, we will omit the explicit notation that indicates the dependence upon x^n. In this case, the above-mentioned examples of information measures will be denoted more simply by $\hat{H}_{\rm nob}(X^d)$, $\hat{H}_{\rm sw}(X^d|Z)$, and $\hat{I}_{\rm csw}(X^d;Z)$, respectively.

2.2.3 LZ Compression and its Properties

The incremental parsing procedure used in the LZ78 algorithm is a sequential method for processing an input sequence x^n drawn from a finite alphabet. At each step, the procedure identifies the shortest substring that has not yet appeared as a complete phrase in the current parsed set except possibly for the final (incomplete) phrase. For instance, applying this parsing method to the sequence

$$x^{15} = \text{abbabaabbaaabaa}$$

yields

a,b,ba,baa,bb,aa,ab,aa.

Let $c(x^n)$ denote the total number of distinct phrases generated by this procedure (in this example, $c(x^{15}) = 8$). Additionally, let $LZ(x^n)$ represent the length in bits of the binary string produced by the LZ78 encoding of x^n. According to Theorem 2 of [12], the following inequality holds:

$$LZ(x^n) \leq [c(x^n) + 1] \log\{2\alpha[c(x^n) + 1]\} \tag{14}$$

which can easily be shown to be further upper bounded by

$$LZ(x^n) \leq c(x^n) \log c(x^n) + n \cdot \epsilon_1(n), \tag{15}$$

where $\epsilon_1(n)$ tends to zero uniformly as $n \to \infty$. In other words, the LZ78 code length for a sequence x^n is upper bounded by an expression whose leading term is $c(x^n) \log c(x^n)$. Remarkably, the very same quantity also appears as the dominant term in a lower bound (see Theorem 1 of [12]) on the shortest code length achievable by any IL finite-state encoder with no more than s states, assuming that $\log(s^2)$ is negligible in comparison to $\log c(x^n)$. More precisely, Theorem 1 in [12] asserts that:

$$\rho_s(x^n) \geq \frac{c(x^n) + s^2}{n} \cdot \log\left(\frac{c(x^n) + s^2}{4s^2}\right) + \frac{2s^2}{n}, \tag{16}$$

which can readily be further lower bounded by

$$\rho_s(x^n) \geq \frac{c(x^n) \log c(x^n)}{n} - \epsilon_2(n, s), \tag{17}$$

where $\epsilon_2(n, s) \to 0$ uniformly as $n \to \infty$ for fixed s. Motivated by this connection, we shall refer to the quantity $c(x^n) \log c(x^n)$ as the unnormalized LZ complexity of x^n The normalized LZ complexity is then defined as

$$\rho_{\text{LZ}}(x^n) \stackrel{\Delta}{=} \frac{c(x^n) \log c(x^n)}{n}, \tag{18}$$

which represents the LZ complexity per input symbol.

3 Bounds on the Average LZ Complexity of k-Blocks

In this section, we derive lower and upper bounds on the average LZ complexity over blocks of length k, that is, on the quantity

$$\frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{\text{LZ}}(x_{ik+1}^{ik+k}) = \frac{1}{n} \sum_{i=0}^{n/k-1} c(x_{ik+1}^{ik+k}) \log c(x_{ik+1}^{ik+k}), \tag{19}$$

where k is a positive integer that divides n. Both the upper bound and the lower bound are given in terms of the empirical entropy $\hat{H}_{\rm csw}(\cdot)$, but to make certain redundancy terms negligibly small, the order of this empirical entropy should be much larger than k in the lower bound and much smaller than k in the upper bound.

The reason for our interest in the average LZ complexity of blocks, rather than in the LZ complexity of the entire sequence, $\rho_{\rm LZ}(x^n)$, is that in any practical application of the LZ78 algorithm, one must reset and start over after each and every block of finite size, as otherwise, the amount of memory and computational effort grows without bound. Also, from the theoretical point of view (see [12, Corollary 2]), the gap between the upper bound and the lower bound to the finite-state complexity of $\boldsymbol{x}$ is closed in the limit of $s \to \infty$, in terms of the double limit

$$\rho(\boldsymbol{x}) = \limsup_{k\to\infty} \limsup_{n\to\infty} \frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{\rm LZ}(x_{ik+1}^{ik+k}). \tag{20}$$

One of the conclusions from our bounds in this section is that, in fact, the outer limit superior over k can be always safely replaced by an ordinary limit, because the sequence

$$\rho_k(\boldsymbol{x}) \stackrel{\Delta}{=} \limsup_{n\to\infty} \frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{\rm LZ}(x_{ik+1}^{ik+k}), \qquad k \in \mathcal{N}, \tag{21}$$

turns out to be convergent thanks to the convergence of the sequence $\{\frac{\hat{H}_{\rm csw}(X^d)[\boldsymbol{x}]}{d}\}_{d\geq 1}$, which plays a role both in the upper bound and in the lower bound, as described above.

3.1 Lower Bound

The following theorem provides our lower bound to the average of the LZ complexities over k-blocks in terms of the cyclic sliding-window empirical entropy.

Theorem 1 *For every positive integer k, every n that is an integer multiple of k, every $x^n \in \mathcal{X}^n$, and every positive integer ℓ,*

$$\frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{LZ}(x_{ik+1}^{ik+k}) \geq \frac{\hat{H}_{csw}(X^\ell)}{\ell} - \Delta_\ell(\alpha^k, \alpha) - \frac{\ell \log \alpha}{n} - \frac{\alpha^\ell}{n} \log \frac{n}{\ell} - \epsilon_1(k), \tag{22}$$

where $\epsilon_1(\cdot)$ is as in (15) and

$$\Delta_\ell(s, \alpha) \stackrel{\Delta}{=} \frac{1}{\ell} \log \left\{ s^2 \cdot \left[1 + \log \left(1 + \frac{\alpha^\ell}{s} \right) \right] \right\}. \tag{23}$$

Since the left-hand side of (22) does not depend on ℓ, in principle, one could maximize the right-hand side over ℓ to obtain the tightest lower bound. But perhaps a more natural point of view is to consider the regime $n \gg \ell \gg k \gg 1$, where the leading term of the lower bound is the empirical entropy term and all other four terms are negligibly small. In particular, taking the limit superior $n \to \infty$, followed by the limit $\ell \to \infty$, and finally, the limit inferior $k \to \infty$, we obtain the following asymptotic inequality as a consequence of eq. (22):

$$\liminf_{k\to\infty} \rho_k(\boldsymbol{x}) \geq \bar{H}_{\mathrm{csw}}[\boldsymbol{x}]. \tag{24}$$

For later use, we point out that the lower bound of Theorem 1 can also be expressed in terms of the empirical entropy associated with non-overlapping blocks, in the following manner:

$$\frac{k}{n}\sum_{i=0}^{n/k-1} \rho_{\mathrm{LZ}}(x_{ik+1}^{ik+k}) \geq \frac{\hat{H}_{\mathrm{nob}}(X^\ell)}{\ell} - \Delta_\ell(\alpha^k, \alpha) - \epsilon_1(k), \tag{25}$$

whose proof is very similar to (and even slightly simpler than) the proof of Theorem 1 below. This will be used in Section 4.

The remaining part of this subsection is devoted to the proof of Theorem 1.

Proof of Theorem 1. We commence by providing a generalized version of Kraft's inequality that applies to any s-state IL encoder. It is similar but somewhat different (and slightly tighter) than the generalized Kraft inequality of [12, Lemma 2].

Lemma 1 *For every IL encoder with s states and every $z \in \mathcal{Z}$,*

$$K(z) \stackrel{\Delta}{=} \sum_{w^\ell \in \mathcal{X}^\ell} 2^{-L[f(z,w^\ell)]} \leq s \cdot \left[1 + \log\left(1 + \frac{\alpha^\ell}{s}\right)\right]. \tag{26}$$

The proof of Lemma 1 is identical to the proof of Lemma 2 of [12] except that since the initial state z is given and fixed, the number k_j of different $\{w^\ell\}$ with $L[f(z, w^\ell)] = j$ cannot exceed $s \cdot 2^j$ (rather than $s^2 2^j$ in [12]), which is the number of combinations of final states and binary output sequences of length j.

Next, observe that

$$s \cdot \left[1 + \log\left(1 + \frac{\alpha^\ell}{s}\right)\right]$$

$$
\begin{aligned}
&\geq \sum_{w^\ell \in \mathcal{X}^\ell} 2^{-L[f(z,w^\ell)]} \\
&= \sum_{w^\ell \in \mathcal{X}^\ell} \hat{P}_{\rm sw}(w^\ell|z) \cdot 2^{-L[f(z,w^\ell)] - \log \hat{P}_{\rm sw}(w^\ell|z)} \\
&\geq \exp_2 \left\{ - \sum_{w^\ell \in \mathcal{X}^\ell} \hat{P}_{\rm sw}(w^\ell|z) \cdot L[f(z,w^\ell)] - \sum_{w^\ell} \hat{P}_{\rm sw}(w^\ell|z) \log \hat{P}_{\rm sw}(w^\ell|z) \right\} \\
&= \exp_2 \left\{ \hat{H}_{\rm sw}(X^\ell|Z=z) - \sum_{w^\ell \in \mathcal{X}^\ell} \hat{P}_{\rm sw}(w^\ell|z) \cdot L[f(z,w^\ell)] \right\}, \qquad (27)
\end{aligned}
$$

where in the second inequality we have used Jensen's inequality and the convexity of the exponential function, $F(u) = 2^u$. This implies that

$$
\sum_{w^\ell \in \mathcal{X}^\ell} \hat{P}_{\rm sw}(w^\ell|z) \cdot L[f(z,w^\ell)] \geq \hat{H}_{\rm sw}(X^\ell|Z=z) - \log \left\{ s \cdot \left[1 + \log \left(1 + \frac{\alpha^\ell}{s} \right) \right] \right\}. \qquad (28)
$$

Averaging both sides w.r.t. $\hat{P}_{\rm sw}(z)$, $z \in \mathcal{Z}$, we end up with

$$
\sum_{(z,w^\ell) \in \mathcal{Z} \times \mathcal{X}^\ell} \hat{P}_{\rm sw}(z,w^\ell) \cdot L[f(z,w^\ell)] \geq \hat{H}_{\rm sw}(X^\ell|Z) - \log \left\{ s \cdot \left[1 + \log \left(1 + \frac{\alpha^\ell}{s} \right) \right] \right\}. \qquad (29)
$$

We next apply this inequality in a chain of inequalities that would lead to a lower bound to $\rho_E(x^n)$. Similarly as in eq. (33) of [12]

$$
\begin{aligned}
\rho_E(x^n) &= \frac{1}{n} \sum_{i=1}^{n} l[f(z_i,x_i)] \\
&= \frac{1}{n\ell} \sum_{i=1}^{n} \ell \cdot l[f(z_i,x_i)] \\
&\geq \frac{1}{n\ell} \sum_{i=0}^{n-\ell} \sum_{j=1}^{\ell} l[f(z_{i+j},x_{i+j})] \\
&= \frac{1}{n\ell} \sum_{i=0}^{n-\ell} L[f(z_{i+1},x_{i+1}^{i+\ell})] \\
&= \frac{1}{\ell} \left(1 - \frac{\ell-1}{n} \right) \cdot \sum_{z,w^\ell} \hat{P}_{\rm sw}(z,w^\ell) \cdot L[f(z,w^\ell)] \\
&\geq \left(1 - \frac{\ell}{n} \right) \cdot \frac{\hat{H}_{\rm sw}(X^\ell|Z)}{\ell} - \frac{1}{\ell} \log \left\{ s \cdot \left[1 + \log \left(1 + \frac{\alpha^\ell}{s} \right) \right] \right\} \\
&= \left(1 - \frac{\ell}{n} \right) \cdot \frac{\hat{H}_{\rm sw}(X^\ell) - \hat{I}_{\rm sw}(Z;X^\ell)}{\ell} - \frac{1}{\ell} \log \left\{ s \cdot \left[1 + \log \left(1 + \frac{\alpha^\ell}{s} \right) \right] \right\} \\
&\geq \left(1 - \frac{\ell}{n} \right) \cdot \frac{\hat{H}_{\rm sw}(X^\ell) - \hat{H}_{\rm sw}(Z)}{\ell} - \frac{1}{\ell} \log \left\{ s \cdot \left[1 + \log \left(1 + \frac{\alpha^\ell}{s} \right) \right] \right\}
\end{aligned}
$$

$$\geq \left(1-\frac{\ell}{n}\right)\cdot\frac{\hat{H}_{\rm sw}(X^\ell)-\log s}{\ell}-\frac{1}{\ell}\log\left\{s\cdot\left[1+\log\left(1+\frac{\alpha^\ell}{s}\right)\right]\right\}$$
$$\geq \frac{\hat{H}_{\rm sw}(X^\ell)}{\ell}-\frac{\ell\log\alpha}{n}-\frac{1}{\ell}\log\left\{s^2\cdot\left[1+\log\left(1+\frac{\alpha^\ell}{s}\right)\right]\right\}. \quad (30)$$

We would now like to modify this lower bound to be given in terms of the empirical entropy $\hat{H}_{\rm csw}(X^\ell)$. It is easy to verify that $|\hat{P}_{\rm csw}(w^\ell)-\hat{P}_{\rm sw}(w^\ell)|\leq\frac{\ell}{n}$ for all $w^\ell\in\mathcal{X}^\ell$, and so, the variational distance between $\hat{P}_{\rm csw}(\cdot)$ and $\hat{P}_{\rm sw}(\cdot)$ cannot exceed $\theta\stackrel{\Delta}{=}\ell\alpha^\ell/n$. Thus, by [1, Lemma 2.7, p. 19],

$$\hat{H}_{\rm sw}(X^\ell)\geq\hat{H}_{\rm csw}(X^\ell)-\frac{\ell\alpha^\ell}{n}\log\frac{n}{\ell}, \quad (31)$$

and so, we have proved that

$$\rho_s(x^n)\geq\frac{\hat{H}_{\rm csw}(X^\ell)}{\ell}-\Delta_\ell(s,\alpha)-\frac{\ell\log\alpha}{n}-\frac{\alpha^\ell}{n}\log\frac{n}{\ell}. \quad (32)$$

Consider now the application of the LZ78 algorithm along blocks of length k, where after each such block the LZ algorithm is restarted independently of previous blocks. Since this is actually a block code of block length k and a block code can be implemented using a finite-state encoder with $s=\alpha^k$ states, then we have:

$$\frac{1}{n}\sum_{i=0}^{n/k-1}LZ(x_{ik+1}^{ik+k}) \geq \rho_{\alpha^k}(x^n)$$
$$\geq \frac{\hat{H}_{\rm csw}(X^\ell)}{\ell}-\Delta_\ell(\alpha^k,\alpha)-\frac{\ell\log\alpha}{n}-\frac{\alpha^\ell}{n}\log\frac{n}{\ell}. \quad (33)$$

On the other hand,

$$\frac{1}{n}\sum_{i=0}^{n/k-1}LZ(x_{ik+1}^{ik+k}) = \frac{k}{n}\sum_{i=0}^{n/k-1}\frac{LZ(x_{ik+1}^{ik+k})}{k}$$
$$\leq \frac{k}{n}\sum_{i=0}^{n/k-1}\rho_{\rm LZ}(x_{ik+1}^{ik+k})+\epsilon_1(k), \quad (34)$$

and so,

$$\frac{k}{n}\sum_{i=0}^{n/k-1}\rho_{\rm LZ}(x_{ik+1}^{ik+k})\geq\frac{\hat{H}_{\rm csw}(X^\ell)}{\ell}-\Delta_\ell(\alpha^k,\alpha)-\frac{\ell\log\alpha}{n}-\frac{\alpha^\ell}{n}\log\frac{n}{\ell}-\epsilon_1(k), \quad (35)$$

thus completing the proof of Theorem 1.

3.2 Upper Bound

Theorem 2 below provides an upper bound to

$$\frac{k}{n}\sum_{i=0}^{n/k-1}\rho_{\mathrm{LZ}}(x_{ik+1}^{ik+k}).$$

Theorem 2 *For every positive integer k, every n that is an integer multiple of k, every $x^n \in \mathcal{X}^n$, and every positive integer m,*

$$\frac{k}{n}\sum_{i=0}^{n/k-1}\rho_{LZ}(x_{ik+1}^{ik+k}) \leq \frac{\hat{H}_{csw}(X^m)}{m} + \frac{1}{m} + \frac{2(m+1)\alpha^{m+1}}{n}\log\frac{n}{m} + \epsilon_2(k,\alpha^{2m}), \tag{36}$$

where $\epsilon_2(\cdot,\cdot)$ is as in eq. (17).

Similarly as in the discussion after Theorem 1, since the left-hand side does not depend on m, in principle, one could minimize the right-hand side over m to obtain the tightest upper bound, but it may be more instructive to consider the regime $n \gg m \gg 1$ and $k \gg m$, where the leading term of the upper bound is the empirical entropy term and all other three terms are negligibly small. In particular, taking the limit superior $n \to \infty$, followed by the limit superior of $k \to \infty$, and finally, the limit $m \to \infty$, we obtain the following asymptotic inequality as a consequence of eq. (36):

$$\limsup_{k\to\infty}\rho_k(\boldsymbol{x}) \leq \bar{H}_{\mathrm{csw}}[\boldsymbol{x}], \tag{37}$$

which together with eq. (24), yields

$$\limsup_{k\to\infty}\rho_k(\boldsymbol{x}) = \liminf_{k\to\infty}\rho_k(\boldsymbol{x}) = \lim_{k\to\infty}\rho_k(\boldsymbol{x}) = \bar{H}_{\mathrm{csw}}[\boldsymbol{x}], \tag{38}$$

in agreement with Theorem 3 of [12], but with the stronger statement that the limit superior over k is actually an ordinary limit, as the sequence $\{\rho_k(\boldsymbol{x})\}_{k\geq 1}$ is convergent.

Similarly as before, the upper bound of Theorem 2 can also be expressed in terms of the empirical entropy associated with non-overlapping blocks, in the following manner:

$$\frac{k}{n}\sum_{i=0}^{n/k-1}\rho_{\mathrm{LZ}}(x_{ik+1}^{ik+k}) \leq \frac{\hat{H}_{\mathrm{nob}}(X^m)}{m} + \frac{1}{m} + \epsilon_2(k,\alpha^{2m}), \tag{39}$$

and once again, the proof is almost identical to the proof of Theorem 2 below. This result too will be used in Section 4.

The remaining part of this subsection is devoted to the proof of Theorem 2.

Proof of Theorem 2. Consider a scenario of compressing x^n by a finite-state encoder with s states that is allowed to vary from one k-block to another. According to eq. (17) (applied to k-blocks), the corresponding compression ratio, which is $\frac{k}{n}\sum_{i=0}^{n/k-1}\rho_s(x_{ik+1}^{ik+k})$, is lower bounded by

$$\frac{k}{n}\sum_{i=0}^{n/k-1}\rho_s(x_{ik+1}^{ik+k}) \geq \frac{k}{n}\sum_{i=0}^{n/k-1}\rho_{\rm LZ}(x_{ik+1}^{ik+k}) - \epsilon_2(k,s). \tag{40}$$

On the other hand, the best time-varying finite-state encoder with $s = \alpha^{2m}$ states (m- positive integer) cannot be worse than the best time-invariant finite-state encoder with the same number of states. Let m divide n and consider block encoding of non-overlapping blocks of length m using a Shannon code with a conditional length function defined by

$$L(x_{im+1}^{im+m}|x_{(i-1)m+1}^{im}) = \left\lceil -\log\left[\prod_{j=1}^{m}Q(x_{im+j}|x_{(i-1)m+j}^{im+j-1})\right]\right\rceil, \qquad i = 0,1,2,\ldots,\frac{n}{m}-1 \tag{41}$$

where for $i=0$, x_1^m is understood to be encoded with fixed arbitrary conditioning on $x_{-(m-1)}^0 \in \mathcal{X}^m$. Next, observe that

$$\begin{aligned}
L(x^n|x_{-(m-1)}^0) &\stackrel{\Delta}{=} \sum_{i=0}^{n/m-1}L(x_{im+1}^{im+m}|x_{(i-1)m+1}^{im}) \\
&= \sum_{i=0}^{n/m-1}\left\lceil -\log\left[\prod_{j=1}^{m}Q(x_{im+j}|x_{(i-1)m+j}^{im+j-1})\right]\right\rceil \\
&\leq -\sum_{i=0}^{n/m-1}\log\left[\prod_{j=1}^{m}Q(x_{im+j}|x_{(i-1)m+j}^{im+j-1})\right] + \frac{n}{m} \\
&= -\sum_{i=0}^{n/m-1}\sum_{j=1}^{m}\log Q(x_{im+j}|x_{(i-1)m+j}^{im+j-1}) + \frac{n}{m} \\
&= -\sum_{i=1}^{n}\log Q(x_i|x_{i-m}^{i-1}) + \frac{n}{m} \\
&= -n\cdot\sum_{w^{m+1}\in\mathcal{X}^{m+1}}\tilde{P}_{\rm sw}(w^{m+1})\log Q(w_{m+1}|w^m) + \frac{n}{m},
\end{aligned} \tag{42}$$

where

$$\tilde{P}_{\rm sw}(w^{m+1}) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{I}\{x_{i-m}^i = w^{m+1}\}, \qquad w^{m+1}\in\mathcal{X}^{m+1}, \tag{43}$$

and in the sequel, we denote the induced entropy by $\tilde{H}_{\rm sw}(\cdot)$. Thus,

$$\frac{k}{n}\sum_{i=0}^{n/k-1}\rho_{\alpha^{2m}}(x_{ik+1}^{ik+k}) \le -\sum_{w^{m+1}\in\mathcal{X}^{m+1}}\tilde{P}_{\rm sw}(w^{m+1})\log Q(w_{m+1}|w^m)+\frac{1}{m}, \tag{44}$$

and since this holds for every $Q(\cdot|\cdot)$, it also holds for the minimizing $Q(\cdot|\cdot)$, which yields

$$\frac{k}{n}\sum_{i=0}^{n/k-1}\rho_{\alpha^{2m}}(x_{ik+1}^{ik+k}) \le \tilde{H}_{\rm sw}(X_{m+1}|X^m)+\frac{1}{m}. \tag{45}$$

At this point, we wish to pass from $\tilde{H}_{\rm sw}(X_{m+1}|X^m)$ to $\hat{H}_{\rm csw}(X_{m+1}|X^m)$, as before. To this end, we first present $\tilde{H}_{\rm sw}(X_{m+1}|X^m)$ as $\tilde{H}_{\rm sw}(X^{m+1})-\tilde{H}_{\rm sw}(X^m)$ and then apply again Lemma 2.7 of [1] to each term separately. Since $|\tilde{P}_{\rm sw}(w^m)-\hat{P}_{\rm csw}(w^m)|\le\frac{m}{n}$ for all $w^m\in\mathcal{X}^m$ then, $\sum_{w^m}|\tilde{P}_{\rm sw}(w^m)-\hat{P}_{\rm csw}(w^m)|\le\frac{m\alpha^m}{n}$, and so, by Lemma 2.7 of [1],

$$|\tilde{H}_{\rm sw}(X^m)-\hat{H}_{\rm csw}(X^m)|\le\frac{m\alpha^m}{n}\log\frac{n}{m} \tag{46}$$

and a similar inequality holds also for $|\tilde{H}_{\rm sw}(X^{m+1})-\hat{H}_{\rm csw}(X^{m+1})|$. It follows that

$$\begin{aligned}
\tilde{H}_{\rm sw}(X_{m+1}|X^m) &= \tilde{H}_{\rm sw}(X^{m+1})-\tilde{H}_{\rm sw}(X^m)\\
&\le \hat{H}_{\rm csw}(X^{m+1})+\frac{(m+1)\alpha^{m+1}}{n}\log\frac{n}{m+1}-\left[\hat{H}_{\rm csw}(X^m)-\frac{m\alpha^m}{n}\log\frac{n}{m}\right]\\
&\le \hat{H}_{\rm csw}(X_{m+1}|X^m)+\frac{2(m+1)\alpha^{m+1}}{n}\log\frac{n}{m},
\end{aligned} \tag{47}$$

and so,

$$\begin{aligned}
\frac{k}{n}\sum_{i=0}^{n/k-1}\rho_{\alpha^{2m}}(x_{ik+1}^{ik+k}) &\le \hat{H}_{\rm csw}(X_{m+1}|X^m)+\frac{1}{m}+\frac{2(m+1)\alpha^{m+1}}{n}\log\frac{n}{m}\\
&\le \frac{1}{m}\sum_{q=1}^{m}\hat{H}_{\rm csw}(X_q|X^{q-1})+\frac{1}{m}+\frac{2(m+1)\alpha^{m+1}}{n}\log\frac{n}{m}\\
&= \frac{\hat{H}_{\rm csw}(X^m)}{m}+\frac{1}{m}+\frac{2(m+1)\alpha^{m+1}}{n}\log\frac{n}{m},
\end{aligned} \tag{48}$$

which yields

$$\frac{k}{n}\sum_{i=0}^{n/k-1}\rho_{\rm LZ}(x_{ik+1}^{ik+k})\le\frac{\hat{H}_{\rm csw}(X^m)}{m}+\frac{1}{m}+\frac{2(m+1)\alpha^{m+1}}{n}\log\frac{n}{m}+\epsilon_2(k,\alpha^{2m}), \tag{49}$$

thus completing the proof of Theorem 2.

4 Chain Rule

Before presenting the chain-rule results, we first need to provide some additional background, which is associated with conditional LZ compression.

4.1 Background on Conditional LZ Compression

In [11], the concept of LZ complexity was extended to account for finite-state lossless compression with side information, leading to the conditional version of LZ complexity. Given sequences x^n and y^n, we apply the incremental parsing procedure of the LZ algorithm to the paired sequence $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$. As previously noted, this procedure ensures that all parsed phrases are distinct, except possibly for the final phrase, which may be incomplete. Let $c(x^n, y^n)$ denote the resulting number of distinct phrases. For instance,[2] if

$$\begin{aligned} x^6 &= 0 \mid 1 \mid 0\ 1 \mid 0\ 1| \\ y^6 &= 0 \mid 1 \mid 0\ 0 \mid 0\ 1| \end{aligned} \tag{50}$$

we have $c(x^6, y^6) = 4$. Let us denote by $c(x^n)$ the resulting number of different phrases of x^n, and denote by $x(l)$ the l-th different x–phrase, $l = 1, 2, \ldots, c(x^n)$. In the running example, $c(x^6) = 3$. Next, let us denote the number of times $x(l)$ appears in the parsing of x^n by $c_l(y^n|x^n)$. Then, obviously, $\sum_{l=1}^{c(x^n)} c_l(y^n|x^n) = c(x^n, y^n)$. In our example, $x(1) = 0$, $x(2) = 1$, $x(3) = 01$, $c_1(y^6|x^6) = c_2(y^6|x^6) = 1$, and $c_3(y^6|x^6) = 2$. Now, the conditional LZ complexity of y^n given x^n is defined as

$$\rho_{LZ}(y^n|x^n) \stackrel{\Delta}{=} \frac{1}{n} \sum_{l=1}^{c(x^n)} c_l(y^n|x^n) \log c_l(y^n|x^n). \tag{51}$$

In [11] it was shown that $\rho_{LZ}(x^n|y^n)$ is the main term of the compression ratio achieved by the conditional version of the LZ algorithm described therein (see also [7]), i.e., the length function, $LZ(x^n|y^n)$, of the coding scheme proposed therein is upper bounded (in parallel to (14)) by

$$LZ(y^n|x^n) \leq n\rho_{LZ}(y^n|x^n) + n\epsilon_3(n), \tag{52}$$

where $\epsilon_3(n)$ is a certain sequence that tends to zero uniformly as $n \to \infty$. On the other hand, analogously to [12, Theorem 1], it was shown in [3], that $\rho_{LZ}(y^n|x^n)$ is also the main term of a

[2] This example is taken from [11].

lower bound to the compression ratio that can be achieved by any finite-state encoder with side information at both ends, provided that the number of states is not too large, similarly as described above for the unconditional version, i.e.,

$$\rho_s(y^n|x^n) \geq \rho_{\rm LZ}(y^n|x^n) - \epsilon_4(n,s), \tag{53}$$

where $\epsilon_4(n,s)$ tends to zero uniformly as $n \to \infty$ for fixed s, and $\rho_s(y^n|x^n)$ is the s-state compressibility of y^n given the side information x^n (available to both encoder and decoder), which is defined in the same manner as the unconditional s-state compressibility, but under an encoder model where the output function f and the next-state function g are fed by both x_i and y_i (in addition to the current state z_i) at each time instant i – see [3] for details.

The results of the previous section can be readily extended to the conditional case. In particular, we will be interested in the relations to the conditional entropy induced by statistics of non-overlapping blocks:

$$\frac{\hat{H}_{\rm nob}(Y^m|X^m)}{m} \leq \frac{q}{n} \sum_{i=0}^{n/q-1} \rho_{\rm LZ}(y_{iq+1}^{iq+q}|x_{iq+1}^{iq+q}) + \epsilon_3(q) + \Delta_m(\alpha^q\beta^q,\beta), \tag{54}$$

and

$$\frac{\hat{H}_{\rm nob}(Y^p|X^p)}{p} \geq \frac{r}{n} \sum_{i=0}^{n/r-1} \rho_{\rm LZ}(y_{ir+1}^{ir+r}|x_{ir+1}^{ir+r}) - \frac{1}{p} - \epsilon_4(r,\alpha^p\beta^p), \tag{55}$$

where m, p, q and r are positive integers, r and q being divisors of n.

4.2 Chain Rule Theorem

Our main result in this section is the following.

Theorem 3 *For every three positive integers k, q and r, every positive integer n that is a multiple of k, q and r, and every $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$, we have:*

1. *Upper bound:*

$$\begin{aligned}\frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{LZ}(x_{ik+1}^{ik+k}, y_{ik+1}^{ik+k}) \quad \leq \quad & \frac{q}{n} \sum_{i=0}^{n/q-1} \Big[\rho_{LZ}(x_{iq+1}^{iq+q}) + \rho_{LZ}(y_{iq+1}^{iq+q}|x_{iq+1}^{iq+q})\Big] + \\ & \Delta_m(\alpha^q,\alpha) + \epsilon_3(q) + \Delta_m(\alpha^q\beta^q,\beta) + \frac{1}{m} + \epsilon_2(k,\alpha^m\beta^m). \end{aligned} \tag{56}$$

2. Lower bound:

$$\frac{k}{n}\sum_{i=0}^{n/k-1}\rho_{LZ}(x_{ik+1}^{ik+k},y_{ik+1}^{ik+k}) \;\geq\; \frac{r}{n}\sum_{i=0}^{n/r-1}\left[\rho_{LZ}(x_{ir+1}^{ir+r})+\rho_{LZ}(y_{ir+1}^{ir+r}|x_{ir+1}^{ir+r})\right]- \\ \frac{2}{p}-\epsilon_2(r,\alpha^p)-\epsilon_4(r,\alpha^p\beta^p)-\Delta_p(\alpha^k\beta^k,\alpha\beta)-\epsilon_1(k). \quad (57)$$

For the upper bound, the integer parameters m and q (q being a divisor of n) are free and can be chosen so as to minimize the right-hand side. In particular, to make all redundancy terms at the second line of eq. (56) small, the regime should be $k \gg m \gg q \gg 1$, which means that we may upper bound the average joint LZ complexity in terms of its decomposition, provided that the block length q of the blocks after the decomposition is very small relative to the original block length, k. Likewise, for the lower bound, the integer parameters p and r (r being a divisor of n) are free and can be chosen so as to maximize the right-hand side. To make all redundancy terms at the second line of eq. (57) small, the regime should be $r \gg p \gg k \gg 1$, which means that we can lower bound the average joint LZ complexity in terms of its decomposition, provided that the block length r of the blocks after the decomposition is very large relative to k. Combining both parts of Theorem 3, the relevant regime is therefore $r \gg p \gg k \gg m \gg q \gg 1$. In view of this, consider eq. (57), take the limit superior of $n \to \infty$, then the limit superior of $r \to \infty$, afterwards the limit of $p \to \infty$ and finally, the limit of $k \to \infty$, to get

$$\rho(\boldsymbol{x},\boldsymbol{y}) \geq \limsup_{r\to\infty}\limsup_{n\to\infty}\frac{r}{n}\sum_{i=0}^{n/r-1}\left[\rho_{\mathrm{LZ}}(x_{ir+1}^{ir+r})+\rho_{\mathrm{LZ}}(y_{ir+1}^{ir+r}|x_{ir+1}^{ir+r})\right]. \quad (58)$$

On the other hand, consider eq. (56), take the limit superior of $n \to \infty$, then the limit of $k \to \infty$, afterwards the limit of $m \to \infty$, and finally the limit inferior of $q \to \infty$, to get

$$\rho(\boldsymbol{x},\boldsymbol{y}) \leq \liminf_{q\to\infty}\limsup_{n\to\infty}\frac{q}{n}\sum_{i=0}^{n/q-1}\left[\rho_{\mathrm{LZ}}(x_{iq+1}^{iq+q})+\rho_{\mathrm{LZ}}(y_{iq+1}^{iq+q}|x_{iq+1}^{iq+q})\right]. \quad (59)$$

It follows that

$$\varrho_k(\boldsymbol{x},\boldsymbol{y}) = \limsup_{n\to\infty}\frac{k}{n}\sum_{i=0}^{n/k-1}\left[\rho_{\mathrm{LZ}}(x_{ik+1}^{ik+k})+\rho_{\mathrm{LZ}}(y_{ik+1}^{ik+k}|x_{ik+1}^{ik+k})\right], \qquad k\in\mathcal{N} \quad (60)$$

is a convergent sequence whose limit $\varrho(\boldsymbol{x},\boldsymbol{y})$ is equal to $\rho(\boldsymbol{x},\boldsymbol{y})$.

Proof of Theorem 3. Let $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ be a given pair of individual sequences. Then, using the results of the previous section, we have the following relations. For the upper bound,

$$\begin{aligned}
& \frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{\mathrm{LZ}}(x_{ik+1}^{ik+k}, y_{ik+1}^{ik+k}) \\
\leq \; & \frac{\hat{H}_{\mathrm{nob}}(X^m, Y^m)}{m} + \frac{1}{m} + \epsilon_2(k, \alpha^m \beta^m) \\
= \; & \frac{\hat{H}_{\mathrm{nob}}(X^m)}{m} + \frac{\hat{H}_{\mathrm{nob}}(Y^m|X^m)}{m} + \frac{1}{m} + \epsilon_2(k, (\alpha\beta)^{2m}) \\
\leq \; & \frac{q}{n} \sum_{i=0}^{n/q-1} \rho_{\mathrm{LZ}}(x_{iq+1}^{iq+q}) + \epsilon_1(q) + \Delta_m(\alpha^q, \alpha) + \\
& \frac{q}{n} \sum_{i=0}^{n/q-1} \rho_{\mathrm{LZ}}(y_{iq+1}^{iq+q}|x_{iq+1}^{iq+q}) + \epsilon_3(q) + \Delta_m(\alpha^q\beta^q, \beta) + \\
& \frac{1}{m} + \epsilon_2(k, \alpha^m \beta^m) \\
= \; & \frac{q}{n} \sum_{i=0}^{n/q-1} \left[\rho_{\mathrm{LZ}}(x_{iq+1}^{iq+q}) + \rho_{\mathrm{LZ}}(y_{iq+1}^{iq+q}|x_{iq+1}^{iq+q})\right] + \\
& \Delta_m(\alpha^q, \alpha) + \epsilon_3(q) + \Delta_m(\alpha^q\beta^q, \beta) + \frac{1}{m} + \epsilon_2(k, \alpha^m \beta^m). \qquad (61)
\end{aligned}$$

For the lower bound,

$$\begin{aligned}
& \frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{\mathrm{LZ}}(x_{ik+1}^{ik+k}, y_{ik+1}^{ik+k}) \\
\geq \; & \frac{\hat{H}_{\mathrm{nob}}(X^p, Y^p)}{p} - \Delta_p(\alpha^k\beta^k, \alpha\beta) - \epsilon_1(k) \\
= \; & \frac{\hat{H}_{\mathrm{nob}}(X^p)}{p} + \frac{\hat{H}_{\mathrm{nob}}(Y^p|X^p)}{p} - \Delta_p(\alpha^k\beta^k, \alpha\beta) - \epsilon_1(k) \\
\geq \; & \frac{r}{n} \sum_{i=0}^{n/r-1} \rho_{\mathrm{LZ}}(x_{ir+1}^{ir+r}) - \frac{1}{p} - \epsilon_2(r, \alpha^p) + \\
& \frac{r}{n} \sum_{i=0}^{n/r-1} \rho_{\mathrm{LZ}}(y_{ir+1}^{ir+r}|x_{ir+1}^{ir+r}) - \frac{1}{p} - \epsilon_4(r, \alpha^p\beta^p) - \\
& \Delta_p(\alpha^k\beta^k, \alpha\beta) - \epsilon_1(k) \\
= \; & \frac{r}{n} \sum_{i=0}^{n/r-1} \left[\rho_{\mathrm{LZ}}(x_{ir+1}^{ir+r}) + \rho_{\mathrm{LZ}}(y_{ir+1}^{ir+r}|x_{ir+1}^{ir+r})\right] - \\
& \frac{2}{p} - \epsilon_2(r, \alpha^p) - \epsilon_4(r, \alpha^p\beta^p) - \Delta_p(\alpha^k\beta^k, \alpha\beta) - \epsilon_1(k). \qquad (62)
\end{aligned}$$

This completes the proof of Theorem 3.

4.3 Comparison to an Earlier Derived Chain Rule

In [4], the following chain-rule theorem was asserted and proved.

Theorem 4 *Define*

$$\begin{aligned}\rho^{+}_{LZ}(x^k, y^k) &= \max\{\rho_{LZ}(x^k, y^k), \rho_{LZ}(x^k) + \rho_{LZ}(y^k|x^k), \\ &\quad \rho_{LZ}(y^k) + \rho_{LZ}(x^k|y^k)\}, \qquad (63)\\ \rho^{-}_{LZ}(x^k, y^k) &= \min\{\rho_{LZ}(x^k, y^k), \rho_{LZ}(x^k) + \rho_{LZ}(y^k|x^k), \\ &\quad \rho_{LZ}(y^k) + \rho_{LZ}(x^k|y^k)\}. \qquad (64)\end{aligned}$$

Given $\boldsymbol{x}$ *and* $\boldsymbol{y}$*, let*

$$\begin{aligned}\rho^{+}(\boldsymbol{x}, \boldsymbol{y}) &= \limsup_{k\to\infty} \limsup_{n\to\infty} \frac{k}{n} \sum_{i=0}^{n/k-1} \rho^{+}_{LZ}(x_{ik+1}^{ik+k}, y_{ik+1}^{ik+k}) \\ \rho^{-}(\boldsymbol{x}, \boldsymbol{y}) &= \limsup_{k\to\infty} \limsup_{n\to\infty} \frac{k}{n} \sum_{i=0}^{n/k-1} \rho^{-}_{LZ}(x_{ik+1}^{ik+k}, y_{ik+1}^{ik+k}).\end{aligned}$$

Then,

$$\rho^{+}(\boldsymbol{x}, \boldsymbol{y}) = \rho^{-}(\boldsymbol{x}, \boldsymbol{y}) = \rho(\boldsymbol{x}, \boldsymbol{y}). \qquad (65)$$

This theorem tells that upon dividing the infinite sequence into non-overlapping k-blocks, then for the infinite sequence pair, it does not matter if on each such block we apply LZ78 compression on $(x_{ik+1}^{ik+k}, y_{ik+1}^{ik+k})$ jointly, or first on x_{ik+1}^{ik+k} and then on y_{ik+1}^{ik+k} given x_{ik+1}^{ik+k}, or vice versa, the ultimate compression ratio will be always the same. However, in contrast to the chain-rule theorem presented here, which applies for finite sequences (with clearly characterized redundancy terms), this theorem of [4] applies to infinite sequences only. Hence, the chain-rule theorem presented here is more refined.

A natural question that may arise at this point is what can be said about the relationship between $\rho(\boldsymbol{x}, \boldsymbol{y})$ and the pair $(\rho(\boldsymbol{x}), \rho(\boldsymbol{y}|\boldsymbol{x}))$ (or $(\rho(\boldsymbol{y}), \rho(\boldsymbol{x}|\boldsymbol{y}))$). On the one hand, it is readily seen that

$$\begin{aligned}&\rho(\boldsymbol{x}, \boldsymbol{y}) \\ =\ &\rho^{-}(\boldsymbol{x}, \boldsymbol{y}) \\ \leq\ &\limsup_{k\to\infty} \limsup_{n\to\infty} \frac{k}{n} \sum_{i=0}^{n/k-1} [\rho_{\mathrm{LZ}}(x_{ik+1}^{ik+k}) + \rho_{\mathrm{LZ}}(y_{ik+1}^{ik+k}|x_{ik+1}^{ik+k})]\end{aligned}$$

$$
\begin{aligned}
&\leq \limsup_{k\to\infty}\limsup_{n\to\infty}\frac{k}{n}\sum_{i=0}^{n/k-1}\rho_{\mathrm{LZ}}(x_{ik+1}^{ik+k})+\\
&\quad \limsup_{k\to\infty}\limsup_{n\to\infty}\frac{k}{n}\sum_{i=0}^{n/k-1}\rho_{\mathrm{LZ}}(y_{ik+1}^{ik+k}|x_{ik+1}^{ik+k})\\
&= \rho(\boldsymbol{x})+\rho(\boldsymbol{y}|\boldsymbol{x}). \qquad (66)
\end{aligned}
$$

However, the reverse inequality, $\rho(\boldsymbol{x},\boldsymbol{y})\geq\rho(\boldsymbol{x})+\rho(\boldsymbol{y}|\boldsymbol{x})$, may not[3] hold true in general, and so, there is no apparent chain rule in that sense. As a counterexample, consider the following. Let $n_0=0$ and $\{n_i,\ i\geq 1\}$ be a sequence of positive integers with the property that for all $i>1$, $n_i\gg\sum_{j=1}^{i-1}n_j$ and consider an infinite binary sequence $\boldsymbol{x}$, defined as follows: For i even and all $n_i+1\leq t\leq n_{i+1}$, $x_t=0$. For i odd and all $n_i+1\leq t\leq n_{i+1}$, x_t is obtained by random coin tossing. Since $n_i\gg\sum_{j=1}^{i-1}n_j$, the compression rate of the last segment always dominates, and so, the compression rate of x^n oscillates forever between 0 and 1, as n grows without bound, which results in a limit superior of $\rho(\boldsymbol{x})=1$ almost surely. Next, let $\boldsymbol{y}$ be defined as follows. For i even and all $n_i+1\leq t\leq n_{i+1}$, y_t is obtained by independent random coin tossing. For i odd and all $n_i+1\leq t\leq n_{i+1}$, we set $y_t=x_t$. Then, $\rho(\boldsymbol{y}|\boldsymbol{x})$ is the limit superior of a sequence of conditional compression rates that oscillates between 0 and 1, and hence $\rho(\boldsymbol{y}|\boldsymbol{x})=1$, implying that $\rho(\boldsymbol{x})+\rho(\boldsymbol{y}|\boldsymbol{x})=2$. On the other hand, when compressing $(\boldsymbol{x},\boldsymbol{y})$ jointly, in each segment the required compression ratio is essentially one bit per symbol pair, (x_t,y_t), in all segments. Therefore, $\rho(\boldsymbol{x},\boldsymbol{y})=1$. It is therefore apparent that the inequality between $\rho(\boldsymbol{x},\boldsymbol{y})$ and $\rho(\boldsymbol{x})+\rho(\boldsymbol{y}|\boldsymbol{x})$ stems mainly from the limit superior operation and the possibility that $\rho(\boldsymbol{x})$ and $\rho(\boldsymbol{y}|\boldsymbol{x})$ may be attained by different subsequences.

References

[1] I. Csiszár and J. Körner, *Information Theory – Coding Theorems for Discrete Memoryless Systems*, Cambridge University Press, New York 2011.

[2] A. N. Kolmogorov, "Logical basis for information theory and probability theory," *IEEE Trans. Inform. Theory*, vol. IT-14, no. 5, pp. 662-664, September 1968.

[3] This corrects a certain mistaken statement in [4].

[3] N. Merhav, "Universal detection of messages via finite–state channels," *IEEE Trans. Inform. Theory*, vol. 46, no. 6, pp. 2242–2246, September 2000.

[4] N. Merhav, "Universal Slepian-Wolf coding for individual sequences," *IEEE Trans. Inform. Theory*, vol. 71, no. 1, pp. 783–796, January 2025.

[5] N. Merhav, "Universal encryption of individual sequences under maximal information leakage," *Entropy*, **2025**, 27(6), 551; May 24, 2025.
`https://doi.org/10.3390/e27060551`

[6] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 471–480, July 1973.

[7] T. Uyematsu and S. Kuzuoka, "Conditional Lempel-Ziv complexity and its application to source coding theorem with side information," *IEICE Trans. Fundamentals*, Vol. E86-A, no. 10, pp. 2615–2617, October 2003.

[8] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inform. Theory*, vol. IT–23, no. 3, pp. 337–343, May 1977.

[9] J. Ziv, "Coding theorems for individual sequences," *IEEE Trans. Inform. Theory*, vol. IT–24, no. 4, pp. 405–412, July 1978.

[10] J. Ziv, "Fixed-rate encoding of individual sequences with side information", *IEEE Transactions on Information Theory*, vol. IT–30, no. 2, pp. 348–352, March 1984.

[11] J. Ziv, "Universal decoding for finite-state channels," *IEEE Trans. Inform. Theory*, vol. IT–31, no. 4, pp. 453–460, July 1985.

[12] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, vol. IT–24, no. 5, pp. 530–536, September 1978.

[13] A. K. Zvonkin and L. A. Levin, "The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms," *Russian Mathematical Surveys*, vol. 25, no. 6, pp. 83–124, 1970.