Universal Slepian-Wolf Coding for Individual Sequences

Neri Merhav

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

Abstract

We establish a coding theorem and a matching converse theorem for separate encodings and joint decoding of individual sequences using finite-state machines. The achievable rate region is characterized in terms of the Lempel-Ziv (LZ) complexities, the conditional LZ complexities and the joint LZ complexity of the two source sequences. An important feature that is needed to this end, which may be interesting on its own right, is a certain asymptotic form of a chain rule for LZ complexities, which we establish in this work. The main emphasis in the achievability scheme is on the universal decoder and its properties. We then show that the achievable rate region is universally attainable by a modified version of Draper's universal incremental Slepian-Wolf (SW) coding scheme, provided that there exists a low-rate reliable feedback link.

Index Terms: Slepian-Wolf coding, Lempel-Ziv algorithm, Lempel-Ziv complexity, finite-state machines, universal decoding.

1 Introduction

The renowned Slepian-Wolf (SW) source coding theorem, first introduced by Slepian and Wolf in 1973 [9], unveils a captivating revelation in the realm of (almost) lossless, fixed-rate compression for memoryless sources in the presence of side information. Remarkably, the theorem establishes that the conditional entropy of the source, given the side information, can be achieved through random binning even if the side information is exclusively available at the decoder, without a necessity for its presence at the encoder. Expanding the horizons of the Slepian-Wolf setting, the theorem is instrumental in characterizing the rate region associated with separate encodings and joint decoding of two correlated memoryless sources. In such scenarios, each coding rate is independently lower bounded by the corresponding conditional entropy, while the rate sum finds its lower bound in the joint entropy. It is imperative to note that in both settings, the joint distribution of the two correlated sources is assumed to be known.

In [1, Problem 13.6, p. 267], Csiszár and Körner considered the case where the joint distribution of the correlated sources is unknown. In this case, the encoders continue to use random binning as before, but the optimal maximum a-posteriori (MAP) decoder is replaced by a universal decoder that seeks a pair of sequences (across the given bins) with minimum joint empirical entropy (see also [5], as well as references therein, for a wider setup of universal source-channel coding and decoding for finite-state sources and finite-state channels with side information). As indicated by Draper [2], the obvious weakness of Csiszár and Körner's universal scheme is that one must commit to fixed coding rates although the source statistics are unknown, and so, there is no mechanism that could adapt the coding rates to the corresponding entropies. Motivated by the will to circumvent this problem, and inspired of earlier works by Shulman [7] and Shulman and Feder [8], Draper proposed a universal, incremental, variable-rate coding scheme that can be implemented provided that a low rate, reliable feedback link is available.

Another perspective of universal source coding is associated with the individual-sequence setting and finite-state machines, as explored by Ziv and Lempel in their celebrated work [13] among some other papers. Indeed, in [11], Ziv studied a scenario of fixed-rate coding with side information where both the source sequence and the side information sequence are individual (deterministic) sequences and where the encoder and decoder are both implementable by finite-state machines. The main

finding in [11] is in establishing and characterizing a notion of fixed-rate conditional complexity as the minimum, almost-lossless compression rate of a source sequence given a side information sequence, and similarly as in classical SW coding, the availability of the side information at the encoder is not necessary in order to achieve this conditional complexity. A year later, in [12], a variable-rate version of the conditional Lempel-Ziv (LZ) complexity was proposed in the completely different context of serving as a universal channel decoding metric for unknown finite-state channels. The utility of this complexity measure in the context of source coding with side information was given further attention later in [4] and [10], but in these works, it was assumed that the side information is available at both ends.

In this work, we consider the framework of SW coding for individual sequences using finite-state encoders. We begin by establishing a coding theorem and a matching converse that together characterize the achievable rate region. Our communication system model is different from that of [11] in several aspects: (i) We consider separate encodings and joint decoding of two individual sequences, as opposed to the narrower problem of encoding a single sequence with the other sequence serving as decoder side information; (ii) Our model for the converse theorem allows variable-rate, finite-state encoding and arbitrary decoding, as opposed to fixed-rate, finite-state encoding and finite-state decoding of [11]; (iii) The relation to the variable-rate coding model in [13] is more apparent; (iv) We establish the variable-rate conditional LZ complexity as the fundamental limit even when the side information is available at the decoder only, and not only when it is available at both ends.

The characterization of the achievable rate region raises an issue which may be interesting on its own right: Recall that in the classical regime of two correlated discrete memoryless sources, X and Y, with joint entropy $H(X,Y)$ and conditional entropies $H(X|Y)$ and $H(Y|X)$, the achievable rate region is given by $\{(R_{\rm x},R_{\rm y}):\ R_{\rm x}\geq H(X|Y),\ R_{\rm y}\geq H(Y|X),R_{\rm x}+R_{\rm y}\geq H(X,Y)\}$, whose corner points are given by $(H(X),H(Y|X))$ and $(H(X|Y),H(Y))$, where the appearance of unconditional marginal entropies, $H(X)$ and $H(Y)$, follows from the chain rule of the entropy, $H(X,Y)=H(X)+H(Y|X)=H(Y)+H(X|Y)$. In the individual-sequence scenario, as we shall see, the achievable region is similar except that $H(X|Y)$, $H(Y|X)$ and $H(X,Y)$ are replaced by the corresponding conditional LZ complexities of the two source sequences and their joint complexity, respectively. However, there is no apparent exact corresponding chain rule that analogously decomposes the

joint LZ complexity of two source sequences, $x^n = (x_1, \ldots, x_n)$ and $y^n = (y_1, \ldots, y_n)$ as the LZ complexity of x^n plus the conditional LZ complexity of y^n given x^n or vice versa. Nonetheless, we will show that at least in a certain asymptotic sense, such as chain rule between the LZ complexities actually applies. This will be instrumental for nailing down the characterization of the corner points of the achievable region in an appealing manner.

In the second part of the paper, we propose a modification of Draper's incremental scheme that is suitable to individual sequences along with their LZ complexities. This will be possible by drawing simple analogies between the various ingredients in Draper's scheme to their corresponding analogues in our individual-sequence setting.

The outline of the remaining part of the paper is as follows. In Section 2, we establish notation conventions, formulate the problem model, define the objectives of this work, and provide some background. In Section 3, we assert and discuss the main coding theorem (Theorem 1) and also establish the asymptotic "chain rule" of the LZ complexity (Theorem 2). Finally, in Section 4, we describe and analyze the universal incremental coding scheme that adjusts the coding rates dynamically. Lengthy proofs are deferred to appendices.

2 Notation, Formulation, Objectives and Background

2.1 Notation

Throughout the paper, random variables will be denoted by capital letters, specific values they may take will be denoted by the corresponding lower case letters, and their alphabets will be denoted by calligraphic letters. Random vectors, their realizations and their alphabets will be denoted, respectively, by capital letters, the corresponding lower case letters, and the corresponding calligraphic letters, all superscripted by their dimension. For example, the random vector $X^n = (X_1, \ldots, X_n)$, (n – positive integer) may take a specific vector value $x^n = (x_1, \ldots, x_n)$ in $\mathcal{X}^n$, the n–th order Cartesian power of the single-letter alphabet $\mathcal{X}$, which will be assumed to have a finite cardinality, α. The notation x_i^j, for $i < j$, will be used to designate the substring $(x_i, x_{i+1}, \ldots, x_j)$. For $i = 1$, the subscript i will be omitted, just like in the notation x^n. Infinite sequences will be denoted using the bold face font, for example, $\boldsymbol{x}$ will designate the sequence $(x_1, x_2, \ldots)$. Similar conventions will apply to other vectors and sequences, such as $y^n \in \mathcal{Y}^n$, y_i^j, and $\boldsymbol{y}$. The single-

letter alphabet $\mathcal{Y}$ will also be assumed to be finite and its cardinality will be denoted by β. The probability of an event $\mathcal{A}$ will be denoted by $\Pr\{\mathcal{A}\}$. Entropies will be denoted using the customary information-theoretic notation, like $H(X)$, $H(Y)$, $H(X,Y)$, $H(X|Y)$, etc., and the same holds for random vectors. The indicator function of an event $\mathcal{A}$ will be denoted by $\mathcal{I}\{A\}$. The notation $[x]_+$ will stand for $\max\{0,x\}$. The cardinality of a finite set $\mathcal{A}$ will be denoted by $|\mathcal{A}|$.

For a given positive integer n and a given ℓ that divides n, the empirical distribution of non-overlapping ℓ-blocks associated with a vector pair $(x^n,y^n)\in\mathcal{X}^n\times\mathcal{Y}^n$, which will be denoted by $\hat{P}$, is the set of relative frequencies

$$\hat{P}(x^\ell,y^\ell)=\frac{\ell}{n}\sum_{i=0}^{n/\ell-1}\mathcal{I}\{(x_{i\ell+1}^{i\ell+\ell},y_{i\ell+1}^{i\ell+\ell})=(x^\ell,y^\ell)\},\qquad (x^\ell,y^\ell)\in\mathcal{X}^\ell\times\mathcal{Y}^\ell. \tag{1}$$

Hereafter, X^ℓ and Y^ℓ will denote auxiliary random vectors of dimension ℓ, jointly distributed according to $\hat{P}$. Accordingly, we will denote by $\hat{H}(X^\ell)$, $\hat{H}(Y^\ell)$, $\hat{H}(X^\ell,Y^\ell)$, $\hat{H}(X^\ell|Y^\ell)$, etc., the various entropies and conditional entropies associated with (X^ℓ,Y^ℓ). In the sequel, we will also divide the pair of n-vectors (x^n,y^n) into n/k non-overlapping segments, each of length k (where k divides n) and within each such segment, $(x_{ik+1}^{ik+k},y_{ik+1}^{ik+k})$, $i=0,1,\ldots,n/k-1$, we define the empirical distribution of non-overlapping ℓ-blocks, $\hat{P}_i$ (assuming that k is divisible by ℓ), and denote the corresponding auxiliary random ℓ-vectors by X_i^ℓ and Y_i^ℓ, respectively. Clearly, $\hat{P}(x^\ell,y^\ell)=\frac{k}{n}\sum_{i=0}^{n/k-1}\hat{P}_i(x^\ell,y^\ell)$.

2.2 Formulation

Let $\boldsymbol{x}=(x_1,x_2,\ldots)$ and $\boldsymbol{y}=(y_1,y_2,\ldots)$ be two individual sequences whose single-letter alphabets, $\mathcal{X}$ and $\mathcal{Y}$, have finite cardinalities, α and β, respectively. Both sequences are to be compressed almost losslessly by separate encoders and jointly decompressed by a central decoder. Each one of encoders is a finite-state encoder defined similarly as in [13], but with a small twist that allows some arbitrarily small distortion (to make the model broad enough to include Slepian-Wolf coding). Also, since the formulation of the setting in this section is for the purpose of the converse bound, it is legitimate to broaden the class of the encoders by allowing them to be "genie-aided" encoders, i.e., letting each one of them to have (sequential) access to the other source as side information. Of course, in the achievability scheme, such an access will not be allowed.

The encoder for $\boldsymbol{x}$, henceforth referred to as the *x-encoder*, is defined by the set

$$E_{\rm x} = (\mathcal{S}, \mathcal{X}, \mathcal{Y}, \mathcal{U}, g_{\rm x}, f_{\rm x}),$$

where $\mathcal{S}$ is the set of states, $\mathcal{X}$ is the input alphabet as mentioned, $\mathcal{Y}$ is the side information alphabet, $\mathcal{U}$ is a finite set of binary output strings, $g_{\rm x} : \mathcal{S} \times \mathcal{X} \times \mathcal{Y} \to \mathcal{S}$ is the next-state function, and $f_{\rm x} : \mathcal{S} \times \mathcal{X} \times \mathcal{Y} \to \mathcal{U}$ is the output function. The members of $\mathcal{U}$ are allowed to be of different lengths, including the empty word λ of length zero. When the input $\boldsymbol{x} = (x_1, x_2, \ldots)$ feeds the encoder $E_{\rm x}$, it produces an output sequence $\boldsymbol{u} = (u_1, u_2, \ldots)$, $u_i \in \mathcal{U}$, $i = 1, 2, \ldots$, while traversing an infinite sequence of states, $\boldsymbol{s} = (s_1, s_2, \ldots)$, $s_i \in \mathcal{S}$, in accordance to

$$u_i = f_{\rm x}(s_i, x_i, y_i) \tag{2}$$

$$s_{i+1} = g_{\rm x}(s_i, x_i, y_i), \quad i = 1, 2, \ldots \tag{3}$$

where s_1 is assumed a fixed member of $\mathcal{S}$. The functions $f_{\rm x}$ and $g_{\rm x}$ are allowed to be "slowly time-varying" in the sense that for some large positive integer k, these functions are piece-wise constant over blocks of length k. In other words, $f_{\rm x}$ and $g_{\rm x}$ are allowed to depend on the running time index i via the quantity $\lfloor i/k \rfloor$. To avoid cumbersome notation, however, we will not indicate this dependence explicitly in the mathematical derivations. Eventually, the parameter k will tend to infinity, which means that the temporal variability is asymptotically slower than that of any time-varying system.

The encoder for $\boldsymbol{y}$, henceforth referred to as the *y-encoder*, is defined exactly in the same manner, except that the roles of the two sources are swapped and accordingly, the subscript "x" is replaced by "y" in all places. Also, $\mathcal{S}$ is replaced by $\mathcal{Z}$, $\mathcal{U}$ is replaced by $\mathcal{V}$, and accordingly, $\boldsymbol{s}$, s_i, $\boldsymbol{u}$ and u_i are substituted by $\boldsymbol{z}$, z_i, $\boldsymbol{v}$ and v_i, respectively. Without an essential loss of generality, the number of states of both $E_{\rm x}$ and $E_{\rm y}$, namely $|\mathcal{S}|$ and $|\mathcal{Z}|$, will be assumed the same, and both will be denoted by q.

As in [13], we adopt the extended notation $f_{\rm x}(s_i, x_i^j, y_i^j)$ and $g_{\rm x}(s_i, x_i^j, y_i^j)$ to denote the output segment u_i^j and final state s_j resulting when the input string x_i^j feeds $E_{\rm x}$, which is at state s_i at time i. Likewise, $g_{\rm y}(z_i, y_i^j, x_i^j)$ and $f_{\rm y}(z_i, y_i^j, x_i^j)$ play analogous roles for the y-encoder.

Let $\epsilon \in (0, 1)$ be a given arbitrarily small number. We assume that $(E_{\rm x}, E_{\rm y})$ together with their joint decoder D form an *ϵ-lossy system*, which is defined as follows: For every $(z_1, s_1) \in \mathcal{Z} \times \mathcal{S}$

and all sufficiently large ℓ, the six-tuple $(z_1, s_1, f_{\rm x}(s_1, x^\ell, y^\ell), f_{\rm y}(z_1, y^\ell, x^\ell), g_{\rm x}(s_1, x^\ell, y^\ell), g_{\rm y}(z_1, y^\ell, x^\ell))$ uniquely determines a decoder output string pair $(\hat{x}^\ell, \hat{y}^\ell)$, whose normalized Hamming distance from (x^ℓ, y^ℓ) does not exceed ϵ, namely, $\frac{1}{\ell}\sum_{i=1}^{\ell}\mathcal{I}\{(\hat{x}_i, \hat{y}_i) \neq (x_i, y_i)\} \leq \epsilon$. In addition, the quadruplet $(y^\ell, s_1, f_{\rm x}(s_1, x^\ell, y^\ell), g_{\rm x}(s_1, x^\ell, y^\ell))$ determines $\hat{x}^\ell$ within normalized Hamming distance ϵ away from x^ℓ, and vice versa: $(x^\ell, z_1, f_{\rm y}(z_1, y^\ell, x^\ell), g_{\rm y}(z_1, y^\ell, x^\ell))$ yields $\hat{y}^\ell$, whose normalized Hamming distance from y^ℓ is at most ϵ.

Let $\mathcal{E}(q, \epsilon)$ denote the class of all ϵ-lossy pairs of finite-state encoders, $(E_{\rm x}, E_{\rm y})$, with no more than q states each. The total number of combinations of states of $E_{\rm x}$ with states of $E_{\rm y}$ is therefore no more than q^2.

2.3 Objectives

Given the vector pair (x^n, y^n) formed by the first n symbol pairs of $(\boldsymbol{x}, \boldsymbol{y})$, let $u^n = (u_1, \ldots, u_n) = f_{\rm x}(s_1, x^n, y^n)$ denote the output of $E_{\rm x}$. We define the compression ratio of x^n by $E_{\rm x}$ as

$$\rho_{E_{\rm x}}(x^n) = \frac{L(u^n)}{n}, \tag{4}$$

where $L(u^n) = \sum_{i=1}^{n} l(u_i)$, $l(u_i)$ being the length (in bits) of u_i, and where it should be kept in mind that for the empty string λ, we set $l(\lambda) = 0$. Likewise, for $v^n = (v_1, \ldots, v_n) = f_{\rm y}(z_1, y^n, x^n)$,

$$\rho_{E_{\rm y}}(y^n) = \frac{L(v^n)}{n}, \tag{5}$$

with $L(v^n) = \sum_{i=1}^{n} l(v_i)$ and with $l(v_i)$ denoting the length of v_i.

For a given encoder pair $(E_{\rm x}, E_{\rm y}) \in \mathcal{E}(q, \epsilon)$, let

$$\mathcal{R}_{E_{\rm x}, E_{\rm y}}(\boldsymbol{x}, \boldsymbol{y}) = \left\{(R_{\rm x}, R_{\rm y}):\ R_{\rm x} \geq \limsup_{n\to\infty}\rho_{E_{\rm x}}(x^n),\ R_{\rm y} \geq \limsup_{n\to\infty}\rho_{E_{\rm y}}(y^n)\right\}. \tag{6}$$

Next, define

$$\mathcal{R}_{q,\epsilon}(\boldsymbol{x}, \boldsymbol{y}) = \bigcup_{(E_{\rm x}, E_{\rm y})\in\mathcal{E}(q,\epsilon)} \mathcal{R}_{E_{\rm x}, E_{\rm y}}(\boldsymbol{x}, \boldsymbol{y}), \tag{7}$$

$$\mathcal{R}_{\epsilon}(\boldsymbol{x}, \boldsymbol{y}) = \bigcup_{q\geq 1} \mathcal{R}_{q,\epsilon}(\boldsymbol{x}, \boldsymbol{y}), \tag{8}$$

and finally,

$$\mathcal{R}(\boldsymbol{x}, \boldsymbol{y}) = \bigcap_{\epsilon>0} \mathcal{R}_{\epsilon}(\boldsymbol{x}, \boldsymbol{y}). \tag{9}$$

These definitions are essentially the two-dimensional counterparts of the s-state compressibility of a single source vector x^n [13, eq. (2)], the asymptotic s-state compressibility of $\boldsymbol{x}$ [13, eq. (3)], and the asymptotic finite-state compressibility of $\boldsymbol{x}$ [13, eq. (4)].

Our first objective is to characterize the set of rate pairs, $\mathcal{R}(\boldsymbol{x}, \boldsymbol{y})$. Our second objective is to propose a universal, incremental variable-rate coding scheme that asymptotically achieves $\mathcal{R}(\boldsymbol{x}, \boldsymbol{y})$ with the aid of a low-rate, reliable feedback channel, that allows adaptation of the coding rates to the compressibilities of the two source sequences. We do that by a simple modification of Draper's scheme for memoryless sources [2].

2.4 Background

To support the exposition of both the converse theorem and the achievability theorem, it is necessary to revisit key terms and details related to the 1978 version of the LZ algorithm, also known as the LZ78 algorithm [13]. The incremental parsing procedure of the LZ78 algorithm is a sequential parsing process applied to the source vector x^k. In this procedure, each new phrase is the shortest string not encountered before as a parsed phrase, except for the potential incompleteness of the last phrase. For instance, the incremental parsing of the vector $x^{15} = \text{abbabaabbaaabaa}$ results in a,b,ba,baa,bb,aa,ab,aa. Let $c(x^k)$ denote the number of phrases in x^k resulting from the incremental parsing procedure (in the above example, $c(x^{15}) = 8$). Furthermore, let $LZ(x^k)$ denote the length of the LZ78 binary compressed code for x^k. According to [13, Theorem 2], the following inequality holds:

$$\begin{aligned}
LZ(x^k) &\leq [c(x^k)+1]\log\{2\alpha[c(x^k)+1]\} \\
&= c(x^k)\log[c(x^k)+1] + c(x^k)\log(2\alpha) + \log\{2\alpha[c(x^k)+1]\} \\
&= c(x^k)\log c(x^k) + c(x^k)\log\left[1+\frac{1}{c(x^k)}\right] + c(x^k)\log(2\alpha) + \log\{2\alpha[c(x^k)+1]\} \\
&\leq c(x^k)\log c(x^k) + \log e + \frac{k(\log\alpha)\log(2\alpha)}{(1-\varepsilon_k)\log k} + \log[2\alpha(k+1)] \\
&\stackrel{\Delta}{=} c(x^k)\log c(x^n) + k\cdot\epsilon(k), \qquad (10)
\end{aligned}$$

where we remind that α is the cardinality of $\mathcal{X}$, and where both ε_k and $\epsilon(k)$ tends to zero as $k \to \infty$. In other words, the LZ code-length for x^k is upper bounded by an expression whose main term is $c(x^k)\log c(x^k)$. On the other hand, $c(x^k)\log c(x^k)$ is also known to be the main term of a lower

bound [13, Theorem 1] to the shortest code-length attainable by any information lossless finite-state encoder with no more than s states, provided that $\log(s^2)$ is very small compared to $\log c(x^k)$. In view of these facts, we henceforth refer to $c(x^k)\log c(x^k)$ as the unnormalized *LZ complexity* of x^k whereas the normalized LZ complexity is defined as

$$\rho_{\rm LZ}(x^k) \stackrel{\Delta}{=} \frac{c(x^k)\log c(x^k)}{k}. \tag{11}$$

A useful inequality, that relates the empirical entropy of non-overlapping ℓ-blocks of x^k (where ℓ divides k) and $\rho_{\rm LZ}(x^k)$ (see, for example, [6, eq. (26)]), is the following:

$$\begin{aligned}\frac{\hat{H}(X^\ell)}{\ell} &\geq \rho_{\rm LZ}(x^k) - \frac{\log[4S^2(\ell)]\log\alpha}{(1-\varepsilon_k)\log k} - \frac{S^2(\ell)\log[4S^2(\ell)]}{k} - \frac{1}{\ell} \\ &\stackrel{\Delta}{=} \rho_{\rm LZ}(x^k) - \Delta_k(\ell),\end{aligned} \tag{12}$$

where

$$S(\ell) = \sum_{i=0}^{\ell-1}\alpha^i = \frac{\alpha^\ell - 1}{\alpha - 1}. \tag{13}$$

It is obtained from the fact that the Shannon code for ℓ-blocks can be implemented using a finite-state encoder with no more than $S(\ell)$ states[1] and therefore it must comply with the lower bound of [13, Theorem 1]. Note that $\lim_{k\to\infty}\Delta_k(\ell) = 1/\ell$ and so, $\lim_{\ell\to\infty}\lim_{k\to\infty}\Delta_k(\ell) = 0$. Clearly, it is possible to let $\ell = \ell(k)$ increase with k slowly enough such that $\Delta_k(\ell(k)) \to 0$ as $k\to\infty$, in particular, $\ell(k)$ should be $o(\log k)$ for that purpose.

In [12], the notion of the LZ complexity was extended to incorporate finite-state lossless compression in the presence of side information, namely, the conditional version of the LZ complexity. Given x^k and y^k, let us apply the incremental parsing procedure of the LZ algorithm to the sequence of pairs $((x_1,y_1),(x_2,y_2),\ldots,(x_k,y_k))$. As mentioned before, according to this procedure, all phrases are distinct with a possible exception of the last phrase, which might be incomplete. Let $c(x^k,y^k)$ denote the number of distinct phrases. For example,[2] if

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

[1] For a block code of length ℓ to be implemented by a finite-state machine, one defines the state at each time instant i to be the contents of the input, starting at the beginning of the current block (at time $\ell\cdot\lfloor i/\ell\rfloor + 1$) and ending at time $i-1$. The number of states for an input alphabet of size α is then $\sum_{i=0}^{\ell-1}\alpha^i = (\alpha^\ell - 1)/(\alpha - 1) < \alpha^\ell$.

[2] The same example appears in [12].

then $c(x^6, y^6) = 4$. Let $c(y^k)$ denote the resulting number of distinct phrases of y^k, and let $y(l)$ denote the l-th distinct y–phrase, $l = 1, 2, \ldots, c(y^k)$. In the above example, $c(y^6) = 3$. Denote by $c_l(x^k|y^k)$ the number of occurrences of $y(l)$ in the parsing of y^k, or equivalently, the number of distinct x-phrases that jointly appear with $y(l)$. Clearly, $\sum_{l=1}^{c(y^k)} c_l(x^k|y^k) = c(x^k, y^k)$. In the above example, $y(1) = 0$, $y(2) = 1$, $y(3) = 01$, $c_1(x^6|y^6) = c_2(x^6|y^6) = 1$, and $c_3(x^6|y^6) = 2$. Now, the conditional LZ complexity of x^k given y^k is defined as

$$\rho_{LZ}(x^k|y^k) \stackrel{\Delta}{=} \frac{1}{k} \sum_{l=1}^{c(y^k)} c_l(x^k|y^k) \log c_l(x^k|y^k). \tag{14}$$

In [12] it was shown that $\rho_{LZ}(x^k|y^k)$ is the main term of the compression ratio achieved by the conditional version of the LZ algorithm described therein (see also [10]), i.e., the length function, $LZ(x^k|y^k)$, of the coding scheme proposed therein is upper bounded (in parallel to (10)) by

$$LZ(x^k|y^k) \leq k\rho_{LZ}(x^k|y^k) + k\hat{\epsilon}(k), \tag{15}$$

where $\hat{\epsilon}(k)$ is a certain sequence that tends to zero as $k \to \infty$. On the other hand, analogously to [13, Theorem 1], it was shown in [4], that $\rho_{LZ}(x^k|y^k)$ is also the main term of a 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.

The inequality (12) also extends to the conditional case as follows (see [4]):

$$\frac{\hat{H}(X^\ell|Y^\ell)}{\ell} \geq \rho_{\rm LZ}(x^k|y^k) - \Delta'_k(\ell), \tag{16}$$

where $\Delta'_k(\ell)$ is the same as $\Delta_k(\ell)$ except that the expression of $S(\ell)$ included therein is redefined as $(\alpha^\ell\beta^\ell - 1)/(\alpha\beta - 1)$ to accommodate the number of states associated with the conditional version of the aforementioned Shannon code applied to ℓ-blocks. By the same token, we also have

$$\frac{\hat{H}(X^\ell, Y^\ell)}{\ell} \geq \rho_{\rm LZ}(x^k, y^k) - \Delta'_k(\ell). \tag{17}$$

We close this section with a comment that although $\rho_{\rm LZ}(y^k)$ and $\rho_{\rm LZ}(x^k|y^k)$ can be thought of as deterministic counterparts of the entropies and conditional entropies [12], [13], to the best knowledge of the author, there is no apparent parallel "chain rule" that explicitly decomposes $\rho_{\rm LZ}(x^k, y^k)$ as

$\rho_{\text{LZ}}(y^k) + \rho_{\text{LZ}}(x^k|y^k)$ or as $\rho_{\text{LZ}}(x^k) + \rho_{\text{LZ}}(y^k|x^k)$. However, we will be able to establish a certain relationship in this spirit at least some asymptotic sense.[3] As described in the Introduction, this will be instrumental in establishing the "corner points" of the achievable rate region.

3 The Coding Theorem

In [13] there is a coding theorem and a matching converse in terms of the long-term average of $\rho_{\text{LZ}}(\cdot)$ applied in successive non-overlapping blocks of the infinite sequence $\boldsymbol{x}$, namely,

$$\rho(\boldsymbol{x}) = \limsup_{k\to\infty} \limsup_{n\to\infty} \frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{\text{LZ}}(x_{ik+1}^{ik+k}), \tag{18}$$

which is a worst-case approach, where the compression ratio is probed at a sequence of block lengths with the worst possible limit.

A similar approach will be executed here too. In particular, denoting

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

$$\rho_k(x^n|y^n) = \frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{\text{LZ}}(x_{ik+1}^{ik+k}|y_{ik+1}^{ik+k}), \tag{20}$$

$$\rho_k(y^n|x^n) = \frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{\text{LZ}}(y_{ik+1}^{ik+k}|x_{ik+1}^{ik+k}), \tag{21}$$

we define the quantities

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = \limsup_{k\to\infty} \limsup_{n\to\infty} \rho_k(x^n, y^n) \tag{22}$$

$$\rho(\boldsymbol{x}|\boldsymbol{y}) = \limsup_{k\to\infty} \limsup_{n\to\infty} \rho_k(x^n|y^n) \tag{23}$$

$$\rho(\boldsymbol{y}|\boldsymbol{x}) = \limsup_{k\to\infty} \limsup_{n\to\infty} \rho_k(y^n|x^n). \tag{24}$$

Theorem 1 *Consider the setting defined in Subsection 2.2 and let $\mathcal{R}(\boldsymbol{x}, \boldsymbol{y})$ be defined as in Subsection 2.3. Then,*

$$\mathcal{R}(\boldsymbol{x}, \boldsymbol{y}) = R(\boldsymbol{x}, \boldsymbol{y}) \stackrel{\Delta}{=} \{(R_x, R_y) : \ R_x \geq \rho(\boldsymbol{x}|\boldsymbol{y}), \ R_y \geq \rho(\boldsymbol{y}|\boldsymbol{x}), \ R_x + R_y \geq \rho(\boldsymbol{x}, \boldsymbol{y})\}. \tag{25}$$

[3] It is interesting to note, in this context, that the Kolmogorov complexity also obeys a parallel chain rule in a certain asymptotic sense, as asserted by the Kolmogorov-Levin theorem [3], [14].

The converse part of Theorem 1, asserting that $\mathcal{R}(\boldsymbol{x},\boldsymbol{y}) \subseteq R(\boldsymbol{x},\boldsymbol{y})$, is proved in Appendix A, and the direct part, asserting that $R(\boldsymbol{x},\boldsymbol{y}) \subseteq \mathcal{R}(\boldsymbol{x},\boldsymbol{y})$, is proved in Appendix B.

A natural question that arises with respect to the coding theorem concerns the corner points of the rate region. If $R_{\mathrm{x}} = \rho(\boldsymbol{x})$ and $R_{\mathrm{x}} + R_{\mathrm{y}} = \rho(\boldsymbol{x},\boldsymbol{y})$, which is one of the corner points at the boundary of the achievable region, then $R_{\mathrm{y}} = \rho(\boldsymbol{x},\boldsymbol{y}) - \rho(\boldsymbol{x})$. In analogy to the traditional probabilistic setting, where $H(X,Y) - H(X) = H(Y|X)$, it is natural to expect that $R_{\mathrm{y}} = \rho(\boldsymbol{y}|\boldsymbol{x})$. This expectation could be met if we can establish a "chain rule", $\rho(\boldsymbol{x},\boldsymbol{y}) = \rho(\boldsymbol{x}) + \rho(\boldsymbol{y}|\boldsymbol{x})$. While there is no known chain rule for the LZ complexities of finite source strings, it turns out that in the asymptotic limit, such a chain rule actually applies in a certain sense. To this end, we assert the following theorem, whose proof appears in Appendix C.

Theorem 2 *Define*

$$\rho^{+}_{LZ}(x^k,y^k) = \max\{\rho_{LZ}(x^k,y^k), \rho_{LZ}(x^k)+\rho_{LZ}(y^k|x^k), \rho_{LZ}(y^k)+\rho_{LZ}(x^k|y^k)\}, \tag{26}$$

$$\rho^{-}_{LZ}(x^k,y^k) = \min\{\rho_{LZ}(x^k,y^k), \rho_{LZ}(x^k)+\rho_{LZ}(y^k|x^k), \rho_{LZ}(y^k)+\rho_{LZ}(x^k|y^k)\}. \tag{27}$$

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

$$\rho^{+}(\boldsymbol{x},\boldsymbol{y}) = \limsup_{k\to\infty}\limsup_{n\to\infty}\frac{k}{n}\sum_{i=0}^{n/k-1}\rho^{+}_{LZ}(x_{ik+1}^{ik+k}, y_{ik+1}^{ik+k}) \tag{28}$$

$$\rho^{-}(\boldsymbol{x},\boldsymbol{y}) = \limsup_{k\to\infty}\limsup_{n\to\infty}\frac{k}{n}\sum_{i=0}^{n/k-1}\rho^{-}_{LZ}(x_{ik+1}^{ik+k}, y_{ik+1}^{ik+k}). \tag{29}$$

Then,

$$\rho^{+}(\boldsymbol{x},\boldsymbol{y}) = \rho^{-}(\boldsymbol{x},\boldsymbol{y}) = \rho(\boldsymbol{x},\boldsymbol{y}). \tag{30}$$

We remark in passing that, as can be seen in the proof of Theorem 2, the three equivalent quantities of eq. (30) are also equal to yet another important well-known quantity, which is the finite-state compressibility of (x^n,y^n), denoted here by $\varrho_\infty(\boldsymbol{x},\boldsymbol{y})$ [13, eq. (4)]. For convenience, we remind here the definition of the finite-state compressibility in a few steps: Let $\varrho_s(x^n,y^n)$ denote the minimum compression ratio achieved by any information lossless s-state (joint) encoder on (x^n,y^n). Next, define $\varrho_s(\boldsymbol{x},\boldsymbol{y}) = \limsup_{n\to\infty}\varrho_s(x^n,y^n)$, and finally, $\varrho_\infty(\boldsymbol{x},\boldsymbol{y}) = \lim_{s\to\infty}\varrho_s(\boldsymbol{x},\boldsymbol{y})$.

To see why the chain rule, $\rho(\boldsymbol{x},\boldsymbol{y}) = \rho(\boldsymbol{x}) + \rho(\boldsymbol{y}|\boldsymbol{x})$, holds true, consider the following argument. First, observe that

$$\begin{aligned}
\rho(\boldsymbol{x},\boldsymbol{y}) &= \rho^{-}(\boldsymbol{x},\boldsymbol{y}) \\
&\leq \limsup_{k\to\infty} \limsup_{n\to\infty} \frac{k}{n} \sum_{i=0}^{n/k-1} [\rho_{\mathrm{LZ}}(x^k) + \rho_{\mathrm{LZ}}(y^k|x^k)] \\
&\leq \limsup_{k\to\infty} \limsup_{n\to\infty} \frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{\mathrm{LZ}}(x^k) + \limsup_{k\to\infty} \limsup_{n\to\infty} \frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{\mathrm{LZ}}(y^k|x^k) \\
&= \rho(\boldsymbol{x}) + \rho(\boldsymbol{y}|\boldsymbol{x}). \qquad (31)
\end{aligned}$$

To establish the reverse inequality, $\rho(\boldsymbol{x},\boldsymbol{y}) \geq \rho(\boldsymbol{x}) + \rho(\boldsymbol{y}|\boldsymbol{x})$ (and hence equality), consider the following: On the one hand, given that $R_{\mathrm{x}} = \rho(\boldsymbol{x})$, then by the direct part of Theorem 1, the rate $R_{\mathrm{y}} = \rho(\boldsymbol{x},\boldsymbol{y}) - \rho(\boldsymbol{x})$ is achievable for the y-encoder, as the point $(\rho(\boldsymbol{x}), \rho(\boldsymbol{x},\boldsymbol{y}) - \rho(\boldsymbol{x}))$ is at (the boundary of) the achievable rate region. On the other hand, given that $R_{\mathrm{x}} = \rho(\boldsymbol{x})$, the source sequence $\boldsymbol{x}$ becomes side information that is available at both ends, and then the compression ratio for $\boldsymbol{y}$ in the presence of $\boldsymbol{x}$ is lower bounded by the conditional LZ complexity, $\rho(\boldsymbol{y}|\boldsymbol{x})$ [4]. Therefore, $\rho(\boldsymbol{x},\boldsymbol{y}) - \rho(\boldsymbol{x}) \geq \rho(\boldsymbol{y}|\boldsymbol{x})$.

Discussion.

Several aspects of Theorem 1 should be highlighted.

1. It is natural to compare our results to those of Ziv in [11]. As mentioned in the Introduction, the class of communication systems allowed here is somewhat broader. We allow variable-rate codes, as opposed to fixed-rate codes of [11]. Also, in our setting there are no limitations on the decoder whereas in [11], a finite-state decoder is assumed. Finally, we consider the complete setting of SW coding, where both sources are compressed, whereas in [11], only one source is compressed and the other source serves as side information. Finally, our coding theorem is more closely related to those of [12] and [13], since it is about a similar form of the conditional LZ complexity. We show that the conditional LZ complexity, $\rho_{\mathrm{LZ}}(\boldsymbol{x}|\boldsymbol{y})$, is operatively meaningful, not only when the side information is available at both ends, but also when it is available at the decoder only. Similarly as in [11], our direct theorem asserts that the expected Hamming distortion is asymptotically vanishing, where the expectation is w.r.t. the randomness of the code in the ensemble. Since the code is re-selected in every block independently, for an infinite sequence pair, $(\boldsymbol{x},\boldsymbol{y})$, the Hamming distortion eventually

vanishes almost surely.

2. As can be seen in the proof of the direct part of Theorem 1, we apply in each block of length k a universal decoder that maximizes the decoding metric,

$$u(x^k, y^k) = \min\{[R_{\rm x} - \rho_{\rm LZ}(x^k|y^k)], [R_{\rm y} - \rho_{\rm LZ}(y^k|x^k)], [R_{\rm x} + R_{\rm y} - \rho_{\rm LZ}(x^k, y^k)]\} \tag{32}$$

among all pairs of vectors that are consistent with the given bin assignments. Note that it is composed of three universal decoding metrics, each one of which handles a different type of error event: (i) error in x^k only, (ii) error in y^k only, and (iii) error in both x^k and y^k. Note that this is different from the universal decoder of Csiszár and Körner for memoryless sources [1, Problem 13.6(b), page 267], which simply minimizes the joint empirical entropy. The reason that the empirical joint entropy handles successfully all three types of errors is associated with the fact that the empirical entropy satisfies the chain rule, $\hat{H}(X,Y) = \hat{H}(X) + \hat{H}(Y|X) = \hat{H}(Y) + \hat{H}(X|Y)$, and so, for errors of types (i) and (ii) the minimum joint entropy decoder is equivalent to minimizing $\hat{H}(X|Y)$ or $\hat{H}(Y|X)$, respectively. Here, on the other hand, this is not the case, because as mentioned before, there is no apparent chain rule for LZ compression ratios for vectors of finite length. Therefore, three different metrics are required.

3. Another interesting observation about our universal decoder is the following. As is well known, in SW decoding each bin functions like a channel code. Since we are talking about universal decoding, it is not surprising to see here universal decoding metrics in the spirit of the maximum mutual information (MMI) decoder [1] or the minimum conditional entropy decoder, or Ziv's 1985 universal decoding metric [12]. What seems to be less trivial is the fact that these universal decoding metrics continue to work well even in the present context of individual sequences, as in contrast to the setting of [12], here there is no finite-state channel that relates $\boldsymbol{y}$ to $\boldsymbol{x}$, and there is no random coding behind the channel inputs.

4. In the direct part, we use a fixed-rate SW code within each k-block. Clearly, it is problematic to use a fixed-rate code since the joint 'statistics' of $\boldsymbol{x}$ and $\boldsymbol{y}$ are not known ahead of time and it is not possible to know in advance the joint LZ complexities and the conditional LZ complexities within

each such block in order to assign coding rates accordingly. However, with very little feedback from the decoder to the encoder, one could construct an adaptive mechanism that in some way 'learns' the joint statistics. One such scheme, which is a modified version of Draper's scheme [2] is proposed in the next section.

4 Incremental SW coding

In [2], Draper proposed and analyzed a universal incremental SW coding scheme for memoryless sources. It turns out that this scheme can be used almost verbatim in the individual-sequence setting considered here, provided that some adjustments are made. Note that our notation here is somewhat different from that of [2].

Draper assumed that the x-encoder (resp. y-encoder) are both connected to fixed-rate noiseless channels (bit pipes), and that the x-encoder (resp. y-encoder) communicates r_{x} (resp. r_{y}) bits per channel use. More precisely, after m channel uses, the x-encoder (resp. y-encoder) has transmitted $\lfloor mr_{\mathrm{x}} \rfloor$ (resp. $\lfloor mr_{\mathrm{y}} \rfloor$) bits over the channel. The proposed coding scheme works in the following stages:

1. The x-encoder (resp. y-encoder) observes the full block, x^k (resp. y^k) and calculates $\rho_{\mathrm{LZ}}(x^k)$ (resp. $\rho_{\mathrm{LZ}}(y^k)$).

2. The x-encoder (resp. y-encoder) communicates the value of $\rho_{\mathrm{LZ}}(x^k)$ (resp. $\rho_{\mathrm{LZ}}(y^k)$) as a header, using approximately $\log n$ bits. Let $\mathcal{T}(x^k) = \{\tilde{x}^k : \rho_{\mathrm{LZ}}(\tilde{x}^k) = \rho_{\mathrm{LZ}}(x^k)\}$ and $\mathcal{T}(y^k) = \{\tilde{y}^k : \rho_{\mathrm{LZ}}(\tilde{y}^k) = \rho_{\mathrm{LZ}}(y^k)\}$ denote the "type classes" of x^k and y^k, respectively. For each possible type class, the respective encoder and decoder agree (ahead of time) on the order of the list of members of that type class. This order is selected independently at random for each type class of each source.

3. The x-encoder (resp. y-encoder) sequentially transmits successive bits of the binary expansion of the location of x^k (resp. y^k) in the shared list. After m channel uses, the decoder has received the first $\lfloor mr_{\mathrm{x}} \rfloor$ (resp. $\lfloor mr_{\mathrm{y}} \rfloor$) of the location of x^k (resp. y^k) in the list. Each incomplete binary expansion corresponds to a bin of sequences whose locations in the list share the same most significant bits. Since the binary expansions are nested, the bins are

nested as well. Referring to the x-encoder (and similarly, for the y-encoder), initially, the bin $\mathcal{B}_0(x^k)$ is the entire type class, $\mathcal{T}(x^k)$. After the first channel use, the bin shrinks to become $\mathcal{B}_1(x^k)$, which is the set of sequences whose position index share the same first $\lfloor r_{\rm x} \rfloor$ bits. After the second channel use, it shrinks further to $\mathcal{B}_2(x^k)$, which is the set with the same $\lfloor 2r_{\rm x} \rfloor$ most significant bits, and so on.

4. After each channel use, the decoder tests all pairs of sequences $(\tilde{x}^k, \tilde{y}^k)$ that are consistent with the corresponding bins currently known to the decoder. Specifically, after m channel uses, it compares an empirical mutual information, $\hat{I}_m(x^k; y^k)$ (to be defined shortly) to a time-varying threshold, θ_m (to be defined shortly as well). As soon as $\hat{I}_m(\tilde{x}^k; \tilde{y}^k) \geq \theta_m$ for some vector pair, the decoder sends an `ACK` to the encoders, which then cease to transmit.

5. If the x-encoder (resp. y-encoder) has already transmitted $k\rho_{\rm LZ}(x^k)$ bits (resp. $k\rho_{\rm LZ}(y^k)$ bits), excluding the header, it ceases to transmit even it has not yet received an `ACK`.

Now, for $m = 1, 2, \ldots$, define

$$\hat{I}_m(x^k; y^k) = \begin{cases} \rho_{\rm LZ}(x^k) + \rho_{\rm LZ}(y^k) - \rho_{\rm LZ}(x^k, y^k) & mr_{\rm x} < LZ(x^k),\ mr_{\rm y} < LZ(y^k) \\ \rho_{\rm LZ}(x^k) - \rho_{\rm LZ}(x^k|y^k) & mr_{\rm x} < LZ(x^k),\ mr_{\rm y} \geq LZ(y^k) \\ \rho_{\rm LZ}(y^k) - \rho_{\rm LZ}(y^k|x^k) & mr_{\rm x} \geq LZ(x^k),\ mr_{\rm y} < LZ(y^k) \end{cases} \tag{33}$$

and

$$\theta_m = \frac{[LZ(x^k) - mr_{\rm x}]_+ + [LZ(y^k) - mr_{\rm y})]_+}{k} + \epsilon, \tag{34}$$

where $\epsilon > 0$ is arbitrarily small.

To analyze the performance of this coding scheme, we proceed similarly as in [2], but with a few twists. We first define the error event after m channel uses:

$$\mathcal{E}_m = \{(\tilde{x}^k, \tilde{y}^k) \neq (x^k, y^k) :\ \rho_{\rm LZ}(\tilde{x}^k) = \rho_{\rm LZ}(x^k),\ \rho_{\rm LZ}(\tilde{y}^k) = \rho_{\rm LZ}(y^k),\ \hat{I}_m(\tilde{x}^k; \tilde{y}^k) \geq \theta_m\}, \tag{35}$$

and three critical values of m:

$$m_{\rm x} = \frac{LZ(x^k)}{r_{\rm x}}, \tag{36}$$

$$m_{\rm y} = \frac{LZ(y^k)}{r_{\rm y}}, \tag{37}$$

$$m_{\rm xy} = \frac{LZ(x^k, y^k)}{r_{\rm x} + r_{\rm y}}. \tag{38}$$

Now, there are three cases, according to the smallest number among $m_{\rm x}$, $m_{\rm y}$, and $m_{\rm xy}$: (i) $m_{\rm xy} < \min\{m_{\rm x}, m_{\rm y}\}$, (ii) $m_{\rm x} < \min\{m_{\rm y}, m_{\rm xy}\}$, and (iii) $m_{\rm y} < \min\{m_{\rm x}, m_{\rm xy}\}$.

Consider case (i) first. As long as $m < m_{\rm xy}$, the probability of error after m channel uses, is upper bounded by

$$\begin{aligned}
\Pr\{\mathcal{E}_m\} &= \sum_{(\tilde{x}^k,\tilde{y}^k)\in\mathcal{E}_m} 2^{-\lfloor mr_{\rm x}\rfloor-\lfloor mr_{\rm y}\rfloor} \\
&\le 4\cdot \sum_{\{(\tilde{x}^k,\tilde{y}^k):\ LZ(\tilde{x}^k,\tilde{y}^k)\le LZ(x^k)+LZ(y^k)-k\theta_m\}} 2^{-m(r_{\rm x}+r_{\rm y})} \\
&\le 4\cdot \sum_{\{(\tilde{x}^k,\tilde{y}^k):\ LZ(\tilde{x}^k,\tilde{y}^k)\le LZ(x^k)+LZ(y^k)-k\theta_m\}} 2^{-m(r_{\rm x}+r_{\rm y})} \\
&\le 8\cdot 2^{LZ(x^k)+LZ(y^k)-k\theta_m}\cdot 2^{-m(r_{\rm x}+r_{\rm y})} \\
&= 8\cdot 2^{-k\epsilon}, \qquad (39)
\end{aligned}$$

where the last step follows from the same argument as in step (c) of eq. (B.10) (see (B.11)). After m exceeds $m_{\rm xy}$, at least the correct pair, (x^k, y^k), certainly exceeds the threshold, but by that time, the two encoders together have transmitted just above $LZ(x^k, y^k)$ bits.

In case (ii), as long as $m < m_{\rm x}$, the probability of error is the same as before. Once m exceeds $m_{\rm x}$ (and both encoder and decoder know when this happens as they both know $LZ(x^k)$ and $r_{\rm x}$), the x-encoder may cease to transmit and the decoder can decode x^k with high reliability by seeking a vector $\tilde{x}^k$ such that: (i) $LZ(\tilde{x}^k)$ agrees with the given $LZ(x^k)$, and (ii) the first $mr_{\rm x}$ bits of the bin index of $\tilde{x}^k$ agree with those that have been received at the decoder from the x-encoder. With high probability there is only one such $\tilde{x}^k$ and it then must by the correct x^k. At this point, only the transmission of the y-encoder continues. Assuming the x^k was decoded correctly, the probability of error after m steps is now upper bounded by:

$$\begin{aligned}
\Pr\{\mathcal{E}_m\} &= \sum_{\{\tilde{y}^k:\ (x^k,\tilde{y}^k)\in\mathcal{E}_m} 2^{-\lfloor mr_{\rm y}\rfloor} \\
&\le 2\cdot \sum_{\{\tilde{y}^k:\ LZ(\tilde{y}^k|x^k)\le LZ(y^k)-k\theta_m\}} 2^{-mr_{\rm y}} \\
&\le 4\cdot 2^{LZ(y^k)-k\theta_m}\cdot 2^{-mr_{\rm y}} \\
&= 4\cdot 2^{-k\epsilon}. \qquad (40)
\end{aligned}$$

As soon as $mr_{\rm y}$ exceeds $LZ(y^k|x^k)$, the correct pair, (x^k, y^k) exceeds the threshold. At this time,

the transmission of the y-encoder stops too and the decoding of y^k can be carried out with x^k in the role of decoder side information. Case (iii) is exactly like case (ii), except that the roles of x^k and y^k are swapped.

Appendix A

Proof of the converse part of Theorem 1.

We first extend the generalized Kraft inequality of [13, Lemma 2] from information lossless encoders to ϵ-lossy encoders. Specifically, we argue that for any given ϵ-lossy encoder pair with q^2 states (i.e., q states of $E_{\rm x}$ times q states of $E_{\rm y}$),

$$\sum_{(x^\ell,y^\ell)\in\mathcal{X}^\ell\times\mathcal{Y}^\ell} 2^{-\{\min_{s\in\mathcal{S}} L[f_{\rm x}(s,x^\ell,y^\ell)]+\min_{z\in\mathcal{Z}} L[f_{\rm y}(z,y^\ell,x^\ell)]\}} \le q^4 B_\ell(\epsilon)\left(1+\log\left[1+\frac{\alpha^\ell\beta^\ell}{q^4 B_\ell(\epsilon)}\right]\right), \quad \text{(A.1)}$$

where

$$B_\ell(\epsilon)=\sum_{j=0}^{\ell\epsilon}\left(\begin{array}{c} l \\ j \end{array}\right)(\alpha\beta-1)^j \quad \text{(A.2)}$$

is the size of the Hamming sphere of radius $\ell\epsilon$ in the space $\mathcal{X}^\ell\times\mathcal{Y}^\ell$ whose size is $\alpha^\ell\beta^\ell$. Using the Chernoff bound, it can be readily seen that

$$B_\ell(\epsilon)\le 2^{\ell Q(\epsilon)} \quad \text{(A.3)}$$

where

$$Q(\epsilon)=\begin{cases} h_2(\epsilon)+\epsilon\log(\alpha\beta-1) & \epsilon<1-\frac{1}{\alpha\beta} \\ \log(\alpha\beta) & 1-\frac{1}{\alpha\beta}\le\epsilon\le 1 \end{cases} \quad \text{(A.4)}$$

and where $h_2(\epsilon)=-\epsilon\log\epsilon-(1-\epsilon)\log(1-\epsilon)$ is the binary entropy function. Since we consider small values of ϵ the second line in the definition of $Q(\epsilon)$ will not be relevant to our derivations. We henceforth denote

$$\delta(\epsilon)\stackrel{\Delta}{=}h_2(\epsilon)+\epsilon\log(\alpha\beta-1). \quad \text{(A.5)}$$

The proof of eq. (A.1) is exactly the same as the proof of [13, Lemma 2], where the only modification needed is that here, the number k_j of (x^ℓ,y^ℓ) for which $\min_{s\in\mathcal{S}} L[f_{\rm x}(s,x^\ell,y^\ell)]+\min_{z\in\mathcal{Z}} L[f_{\rm y}(z,y^\ell,x^\ell)]=j$ is upper bounded by $q^4B_\ell(\epsilon)2^j$, as follows from the postulate that the encoders form an ϵ-lossy system. It follows that the total description length at the outputs of the encoders is lower bounded

as follows.

$$
\begin{aligned}
n(R_{\mathrm{x}}+R_{\mathrm{y}}) &\geq \sum_{t=1}^{n}\{L[f_{\mathrm{x}}(s_t,x_t,y_t)]+L[f_{\mathrm{y}}(z_t,y_t,x_t)]\} \\
&= \sum_{i=0}^{n/k-1}\sum_{m=0}^{k/\ell-1}\sum_{j=1}^{\ell}\{L[f_{\mathrm{x}}(s_i,x_{ik+m\ell+j},y_{ik+m\ell+j}]+L[f_{\mathrm{y}}(z_i,y_{ik+m\ell+j},x_{ik+m\ell+j}]\} \\
&= \sum_{i=0}^{n/k-1}\sum_{m=0}^{k/\ell-1}\{L[f_{\mathrm{x}}(s_i,x_{ik+m\ell+1}^{ik+m\ell+\ell},y_{ik+m\ell+1}^{ik+m\ell+\ell})]+L[f_{\mathrm{y}}(z_i,y_{ik+m\ell+1}^{ik+m\ell+\ell},x_{ik+m\ell+1}^{ik+m\ell+\ell})]\} \\
&\geq \sum_{i=0}^{n/k-1}\sum_{m=0}^{k/\ell-1}\{\min_{s\in\mathcal{S}}L[f_{\mathrm{x}}(s,x_{ik+m\ell+1}^{ik+m\ell+\ell},y_{ik+m\ell+1}^{ik+m\ell+\ell})]+\min_{z\in\mathcal{Z}}L[f_{\mathrm{y}}(z,y_{ik+m\ell+1}^{ik+m\ell+\ell},x_{ik+m\ell+1}^{ik+m\ell+\ell})]\} \\
&= \sum_{i=0}^{n/k-1}\frac{k}{\ell}\sum_{(x^\ell,y^\ell)\in\mathcal{X}^\ell\times\mathcal{Y}^\ell}\hat{P}_i(x^\ell,y^\ell)\cdot[\min_{s\in\mathcal{S}}L[f_{\mathrm{x}}(s,x^\ell,y^\ell)]+\min_{z\in\mathcal{Z}}L[f_{\mathrm{y}}(z,y^\ell,x^\ell)]. \qquad \text{(A.6)}
\end{aligned}
$$

Now, according to the generalized Kraft inequality,

$$
\begin{aligned}
& q^4B_\ell(\epsilon)\left(1+\log\left[1+\frac{(\alpha\beta)^\ell}{q^4B_\ell(\epsilon)}\right]\right) \\
\geq & \sum_{(x^\ell,y^\ell)\in\mathcal{X}^\ell\times\mathcal{Y}^\ell}\exp_2\{-[\min_{s\in\mathcal{S}}L[f_{\mathrm{x}}(s,x^\ell,y^\ell)]+\min_{z\in\mathcal{Z}}L[f_{\mathrm{y}}(z,y^\ell,x^\ell)]\} \\
= & \sum_{(x^\ell,y^\ell)\in\mathcal{X}^\ell\times\mathcal{Y}^\ell}\hat{P}_i(x^\ell,y^\ell)\cdot\exp_2\{-[\min_{s\in\mathcal{S}}L[f_{\mathrm{x}}(s,x^\ell,y^\ell)]+\min_{z\in\mathcal{Z}}L[f_{\mathrm{y}}(z,y^\ell,x^\ell)]-\log\hat{P}_i(x^\ell,y^\ell)\} \\
\geq & \exp_2\Bigg\{-\sum_{(x^\ell,y^\ell)\in\mathcal{X}^\ell\times\mathcal{Y}^\ell}\hat{P}_i(x^\ell,y^\ell)\cdot\Bigg(\min_{s\in\mathcal{S}}L[f_{\mathrm{x}}(s,x^\ell,y^\ell)]+ \\
& \min_{z\in\mathcal{Z}}L[f_{\mathrm{y}}(z,y^\ell,x^\ell)]\Bigg)+\hat{H}(X_i^\ell,Y_i^\ell)\Bigg\}. \qquad \text{(A.7)}
\end{aligned}
$$

Taking the base 2 logarithms of both sides, this yields

$$
\begin{aligned}
& \log\left\{q^4B_\ell(\epsilon)\left(1+\log\left[1+\frac{(\alpha\beta)^\ell}{q^4B_\ell(\epsilon)}\right]\right)\right\} \\
\geq & \hat{H}(X_i^\ell,Y_i^\ell)-\sum_{(x^\ell,y^\ell)\in\mathcal{X}^\ell\times\mathcal{Y}^\ell}\hat{P}_i(x^\ell,y^\ell)\cdot[\min_{s\in\mathcal{S}}L[f_{\mathrm{x}}(s,x^\ell,y^\ell)]+\min_{z\in\mathcal{Z}}L[f_{\mathrm{y}}(z,y^\ell,x^\ell)], \qquad \text{(A.8)}
\end{aligned}
$$

implying that

$$
\begin{aligned}
& R_{\mathrm{x}} + R_{\mathrm{y}} \\
\geq\ & \frac{k}{n} \sum_{i=0}^{n/k-1} \frac{1}{\ell} \sum_{(x^\ell, y^\ell) \in \mathcal{X}^\ell \times \mathcal{Y}^\ell} \hat{P}_i(x^\ell, y^\ell) \cdot \left(\min_{s \in \mathcal{S}} L[f_{\mathrm{x}}(s, x^\ell, y^\ell)] + \min_{z \in \mathcal{Z}} L[f_{\mathrm{y}}(z, y^\ell, x^\ell)] \right) \\
\geq\ & \frac{k}{n} \sum_{i=0}^{n/k-1} \frac{\hat{H}(X_i^\ell, Y_i^\ell)}{\ell} - \frac{1}{\ell} \log \left\{ q^4 B_\ell(\epsilon) \left(1 + \log \left[1 + \frac{(\alpha\beta)^\ell}{q^4 B_\ell(\epsilon)} \right] \right) \right\} \\
\geq\ & \frac{k}{n} \sum_{i=0}^{n/k-1} \frac{\hat{H}(X_i^\ell, Y_i^\ell)}{\ell} - \frac{1}{\ell} \log \left\{ q^4 \left(1 + \log \left[1 + \frac{(\alpha\beta)^\ell}{q^4} \right] \right) \right\} - \delta(\epsilon) \\
\geq\ & \frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{\mathrm{LZ}}(x_{ik+1}^{ik+k}, y_{ik+1}^{ik+k}) - \Delta_k'(\ell) - \frac{1}{\ell} \log \left\{ q^4 \left(1 + \log \left[1 + \frac{(\alpha\beta)^\ell}{q^4} \right] \right) \right\} - \delta(\epsilon). \quad \text{(A.9)}
\end{aligned}
$$

In the same manner, we can derive a generalized Kraft inequality for the x-encoder when y^ℓ is fixed:

$$
\sum_{x^\ell \in \mathcal{X}^\ell} 2^{-\min_{s \in \mathcal{S}} L[f_{\mathrm{x}}(s, x^\ell, y^\ell)]} \leq q^2 B_\ell(\epsilon) \left(1 + \log \left[1 + \frac{\alpha^\ell}{q^2 B_\ell(\epsilon)} \right] \right), \quad \text{(A.10)}
$$

and likewise, vice versa:

$$
\sum_{y^\ell \in \mathcal{Y}^\ell} 2^{-\min_{z \in \mathcal{Z}} L[f_{\mathrm{y}}(z, y^\ell, x^\ell)]} \leq q^2 B_\ell(\epsilon) \left(1 + \log \left[1 + \frac{\beta^\ell}{q^2 B_\ell(\epsilon)} \right] \right). \quad \text{(A.11)}
$$

Using the same method as above, we arrive at the following individual rate bounds:

$$
R_{\mathrm{x}} \geq \frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{\mathrm{LZ}}(x_{ik+1}^{ik+k} | y_{ik+1}^{ik+k}) - \Delta_k'(\ell) - \frac{1}{\ell} \log \left\{ q^2 \left(1 + \log \left[1 + \frac{\alpha^\ell}{q^2} \right] \right) \right\} - \delta(\epsilon), \quad \text{(A.12)}
$$

and

$$
R_{\mathrm{y}} \geq \frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{\mathrm{LZ}}(y_{ik+1}^{ik+k} | x_{ik+1}^{ik+k}) - \Delta_k'(\ell) - \frac{1}{\ell} \log \left\{ q^2 \left(1 + \log \left[1 + \frac{\beta^\ell}{q^2} \right] \right) \right\} - \delta(\epsilon). \quad \text{(A.13)}
$$

Taking the limit superior of $n \to \infty$, followed by the limit superior of $k \to \infty$, followed in turn by the limit of $\ell \to \infty$, and finally, the limit of $\epsilon \downarrow 0$, in eqs. (A.9), (A.12), and (A.13), we obtain

$$
\begin{aligned}
R_{\mathrm{x}} + R_{\mathrm{y}} &\geq \rho(\boldsymbol{x}, \boldsymbol{y}) && \text{(A.14)} \\
R_{\mathrm{x}} &\geq \rho(\boldsymbol{x} | \boldsymbol{y}) && \text{(A.15)} \\
R_{\mathrm{y}} &\geq \rho(\boldsymbol{y} | \boldsymbol{x}). && \text{(A.16)}
\end{aligned}
$$

This completes the proof of the converse part of Theorem 1.

Appendix B

Proof of the direct part of Theorem 1.

Consider the partition of the source sequence pair, (x^n, y^n), into n/k non-overlapping blocks of length k, $(x_{ik+1}^{ik+k}, y_{ik+1}^{ik+k})$, $0, 1, 2, \ldots, n/k-1$. Select an arbitrarily small $\epsilon_0 > 0$, and for each i, let us select select a pair of coding rates, $(R_{\mathrm{x}}^i, R_{\mathrm{y}}^i)$ such that

$$R_{\mathrm{x}}^i \geq \rho_{\mathrm{LZ}}(x_{ik+1}^{ik+k}|y_{ik+1}^{ik+k}) + \epsilon_0 \tag{B.1}$$

$$R_{\mathrm{y}}^i \geq \rho_{\mathrm{LZ}}(y_{ik+1}^{ik+k}|x_{ik+1}^{ik+k}) + \epsilon_0 \tag{B.2}$$

$$R_{\mathrm{x}}^i + R_{\mathrm{y}}^i \geq \rho_{\mathrm{LZ}}(x_{ik+1}^{ik+k}, y_{ik+1}^{ik+k}) + \epsilon_0, \tag{B.3}$$

so that $R_{\mathrm{x}} = \frac{k}{n}\sum_{i=0}^{n/k-1} R_{\mathrm{x}}^i$ and $R_{\mathrm{y}} = \frac{k}{n}\sum_{i=0}^{n/k-1} R_{\mathrm{x}}^i$ satisfy

$$R_{\mathrm{x}} \geq \rho_k(x^n|y^n) + \epsilon_0 \tag{B.4}$$

$$R_{\mathrm{y}} \geq \rho_k(y^n|x^n) + \epsilon_0 \tag{B.5}$$

$$R_{\mathrm{x}} + R_{\mathrm{y}} \geq \rho_k(x^n, y^n) + \epsilon_0, \tag{B.6}$$

and then, in the limit of large n and k, $(R_{\mathrm{x}}, R_{\mathrm{y}})$ is in $R(\boldsymbol{x}, \boldsymbol{y})$.

Now, for $i = 0, 1, 2, \ldots$, let both x_{ik+1}^{ik+k} and y_{ik+1}^{ik+k} be compressed separately by random binning encoders with block length k, ϕ_{x}^i and ϕ_{y}^i at rates R_{x}^i and R_{y}^i, respectively. This is to say that every $x^k \in \mathcal{X}^k$ (resp. $y^k \in \mathcal{Y}^k$), is mapped into an index $\phi_{\mathrm{x}}^i(x^k)$ (resp. $\phi_{\mathrm{y}}^i(y^k)$) that is selected independently at random across the range $\{0, 1, 2, \ldots, 2^{kR_{\mathrm{x}}^i} - 1\}$ (resp. $\{0, 1, 2, \ldots, 2^{kR_{\mathrm{y}}^i} - 1\}$) under the uniform distribution (independently for every i). Clearly, a block code of length k for x^k (resp. y^k) can be implemented using a finite-state machine with no more than α^k (resp. β^k) states. Note also that this pair of encoders belongs to the class of slowly time-varying encoders, which are piece-wise constant along segments of length k, as described in Subsection 2.2.

Throughout the remaining part of the proof we analyze the probability of error within each block of length k and show that it tends to zero as $k \to \infty$, and so, the expected Hamming distance between the decoded sources and the input sources vanish in the long run. For the sake of notational simplicity, we henceforth avoid the indexing by i. In other words, with a slight abuse of notation, we replace x_{ik+1}^{ik+k}, y_{ik+1}^{ik+k}, ϕ_{x}^i, ϕ_{y}^i, R_{x}^i and R_{y}^i by x^k, y^k, ϕ_{x}, ϕ_{y}, R_{x}, and R_{y}, respectively. Consider

now the following decoder that maps the pair $(\phi_{\rm x}(x^k),\phi_{\rm y}(y^k))$ into the decoded estimates of the sources vectors, $(\hat{x}^k,\hat{y}^k)$:

$$(\hat{x}^k,\hat{y}^k)=\arg\max_{\{(\tilde{x}^k,\tilde{y}^k):\ \phi_{\rm x}(\tilde{x}^k)=\phi_{\rm x}(x^k),\ \phi_{\rm y}(\tilde{y}^k)=\phi_{\rm y}(y^k)\}}u(\tilde{x}^k,\tilde{y}^k), \tag{B.7}$$

where

$$u(\tilde{x}^k,\tilde{y}^k)=\min\{[R_{\rm x}-\rho_{\rm LZ}(\tilde{x}^k|\tilde{y}^k)],\ [R_{\rm y}-\rho_{\rm LZ}(\tilde{y}^k|\tilde{x}^k)],\ [R_{\rm x}+R_{\rm y}-\rho_{\rm LZ}(\tilde{x}^k,\tilde{y}^k)]\}. \tag{B.8}$$

Observe that $u(x^k,y^k)\geq\epsilon_0$ by (B.1)-(B.3 by (B.1)-(B.3)). The error event can be presented as the disjoint union of three types of error events: The first type of error is when $\tilde{y}^k=y^k$ and only $\tilde{x}^k\neq\tilde{x}^k$, the second type is the other way around, and the third type is when both $\tilde{x}^k\neq x^k$ and $\tilde{y}^k\neq y^k$. Accordingly,

$$P_{\rm e}(x^k,y^k)=P_{\rm e1}(x^k,y^k)+P_{\rm e2}(x^k,y^k)+P_{\rm e3}(x^k,y^k), \tag{B.9}$$

where

$$\begin{aligned}
P_{\rm e1}(x^k,y^k) &= \sum_{\{\tilde{x}^k:\ u(\tilde{x}^k,y^k)\geq u(x^k,y^k)\}}\Pr\{\phi_{\rm x}(\tilde{x}^k)=\phi_{\rm x}(x^k)\}\\
&\overset{\rm(a)}{\leq} \sum_{\{\tilde{x}^k:\ kR_{\rm x}-k\rho_{\rm LZ}(\tilde{x}^k|y^k)\geq ku(x^k,y^k)\}}2^{-kR_{\rm x}}\\
&\overset{\rm(b)}{\leq} \sum_{\{\tilde{x}^k:\ LZ(\tilde{x}^k|y^k)\leq kR_{\rm x}-ku(x^k,y^k)+k\hat{\epsilon}(k)\}}2^{-kR_{\rm x}}\\
&\overset{\rm(c)}{\leq} 2\cdot 2^{kR_{\rm x}-ku(x^k,y^k)+k\hat{\epsilon}(k)}\cdot 2^{-kR_{\rm x}}\\
&= 2\cdot 2^{-k[u(x^k,y^k)-\hat{\epsilon}(k)]}\\
&\overset{\rm(d)}{\leq} 2^{-k[\epsilon_0-\hat{\epsilon}(k)-1/k]},
\end{aligned} \tag{B.10}$$

where (a) is since $R_{\rm x}-\rho_{\rm LZ}(\tilde{x}^k|y^k)\geq u(\tilde{x}^k,y^k)$ by definition of $u(\cdot,\cdot)$, (b) stems from (15), and (c) is based on the following consideration: Since $LZ(x^k|y^k)$ is a length function of a lossless code, the size of the set $\{x^k:\ LZ(x^k|y^k)=l\}$ cannot exceed 2^ℓ, and so, for any positive integer L,

$$\left|\{x^k:LZ(x^k|y^k)\leq L\}\right|=\sum_{l=1}^{L}\left|\{x^k:LZ(x^k|y^k)=l\}\right|\leq\sum_{l=1}^{L}2^l=2^{L+1}-1<2\cdot 2^L. \tag{B.11}$$

Finally, (d) holds because $u(x^k,y^k)\geq\epsilon_0$, as observed above. Clearly, $P_{\rm e2}(x^k,y^k)$ is treated exactly in the same way, except that the roles of x^k and y^k are swapped. Therefore,

$$P_{\rm e2}(x^k,y^k)\leq 2^{-k[\epsilon_0-\hat{\epsilon}(k)-1/k]} \tag{B.12}$$

as well. Finally,

$$\begin{aligned} P_{\mathrm{e}3}(x^k,y^k) &= \sum_{\{(\tilde{x}^k,\tilde{y}^k):\ u(\tilde{x}^k,\tilde{y}^k)\geq u(x^k,y^k)\}} \Pr\{\phi_{\mathrm{x}}(\tilde{x}^k)=\phi_{\mathrm{x}}(x^k),\ \phi_{\mathrm{y}}(\tilde{y}^k)=\phi_{\mathrm{y}}(y^k)\} \\ &\leq \sum_{\{(\tilde{x}^k,\tilde{y}^k):\ R_{\mathrm{x}}+R_{\mathrm{y}}-\rho_{\mathrm{LZ}}(\tilde{x}^k,\tilde{y}^k)\geq u(x^k,y^k)\}} 2^{-k(R_{\mathrm{x}}+R_{\mathrm{y}})} \\ &= \sum_{\{(\tilde{x}^k,\tilde{y}^k):\ LZ(\tilde{x}^k,\tilde{y}^k)\leq k(R_{\mathrm{x}}+R_{\mathrm{y}})-ku(x^k,y^k)+k\epsilon(k)\}} 2^{-k(R_{\mathrm{x}}+R_{\mathrm{y}})} \\ &< 2\cdot 2^{k(R_{\mathrm{x}}+R_{\mathrm{y}})-ku(x^k,y^k)+k\epsilon(k)}\cdot 2^{-k(R_{\mathrm{x}}+R_{\mathrm{y}})} \\ &\leq 2^{-k[\epsilon_0-\epsilon(k)-1/k]}, \end{aligned} \tag{B.13}$$

and so, overall, $P_{\mathrm{e}}(x^k,y^k)$ tends to zero as $k\to\infty$. This completes the proof of the direct part of Theorem 1.

Appendix C

Since the inequality $\rho^+(\boldsymbol{x},\boldsymbol{y})\geq\rho^-(\boldsymbol{x},\boldsymbol{y})$ is obvious, it is enough to prove the reverse inequality, $\rho^+(\boldsymbol{x},\boldsymbol{y})\leq\rho^-(\boldsymbol{x},\boldsymbol{y})$. For a given n, consider the following encoding of (x^n,y^n) in blocks of length k, where k is assumed to divide n. For each k-block, $(x_{ik+1}^{ik+k},y_{ik+1}^{ik+k})$, $i=0,1,\ldots,n/k-1$, the encoder compares the length functions of three compression schemes:

1. Scheme A applies LZ compression of the sequence of pairs $(x_{ik+1},y_{ik+1}),\ldots(x_{ik+k},y_{ik+k})$ using $LZ(x_{ik+1}^{ik+k},y_{ik+1}^{ik+k})$ bits.

2. Scheme B first compresses x_{ik+1}^{ik+k} to $LZ(x_{ik+1}^{ik+k})$ bits and then compresses y_{ik+1}^{ik+k} into $LZ(y_{ik+1}^{ik+k}|x_{ik+1}^{ik+k})$ bits by utilizing x_{ik+1}^{ik+k} as side information available at both ends.

3. Scheme C is the same as Scheme B, except that the roles of x_{ik+1}^{ik+k} and y_{ik+1}^{ik+k} are interchanged.

The encoder selects the shortest code among those of schemes A, B and C, and adds a header of two flag bits to indicate to the decoder which one of the three schemes was chosen. The overall coding rate is therefore

$$\frac{1}{n}\sum_{i=0}^{n/k-1}[k\rho_{\mathrm{LZ}}^-(x_{ik+1}^{ik+k},y_{ik+1}^{ik+k})+2]=\frac{k}{n}\sum_{i=0}^{n/k-1}\rho_{\mathrm{LZ}}^-(x_{ik+1}^{ik+k},y_{ik+1}^{ik+k})+\frac{2}{k}. \tag{C.1}$$

This is a block code of length k, and as such, it can be implemented by a finite-state machine with no more than

$$s = \sum_{i=0}^{k-1} \alpha^i \beta^i = \frac{\alpha^k \beta^k - 1}{\alpha\beta - 1} < \alpha^k \beta^k \tag{C.2}$$

states (see footnote no. 1). Therefore,

$$\varrho_{\alpha^k\beta^k}(x^n, y^n) \le \frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{\rm LZ}^-(x_{ik+1}^{ik+k}, y_{ik+1}^{ik+k}) + \frac{2}{k}, \tag{C.3}$$

where $\varrho_s(x^n, y^n)$ and its corresponding limits were defined after Theorem 2. Consequently,

$$\varrho_{\alpha^k\beta^k}(\boldsymbol{x}, \boldsymbol{y}) = \limsup_{n\to\infty} \varrho_{\alpha^k\beta^k}(x^n, y^n) \le \limsup_{n\to\infty} \frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{\rm LZ}^-(x_{ik+1}^{ik+k}, y_{ik+1}^{ik+k}) + \frac{2}{k}. \tag{C.4}$$

Upon taking the limit superior of $k \to \infty$, we get

$$\varrho_\infty(\boldsymbol{x}, \boldsymbol{y}) = \limsup_{k\to\infty} \varrho_{\alpha^k\beta^k}(\boldsymbol{x}, \boldsymbol{y}) = \lim_{k\to\infty} \varrho_{\alpha^k\beta^k}(\boldsymbol{x}, \boldsymbol{y}) \le \rho^-(\boldsymbol{x}, \boldsymbol{y}). \tag{C.5}$$

On the other hand, as a lower bound to the lossless compression ratio for the ith k-block ($i = 0, 1, \ldots, n/k - 1$)) by an s-state encoder, we can apply the second to the last line of eq. (A.9) with $\epsilon = 0$ to obtain:

$$\begin{aligned} \frac{L_i}{k} &\ge \frac{\hat{H}_\ell(X_i^\ell, Y_i^\ell)}{\ell} - \delta_s(\ell) \\ &\ge \rho_{\rm LZ}(x_{ik+1}^{ik+k}, y_{ik+1}^{ik+k}) - \delta_s(\ell) - \Delta_k'(\ell) \end{aligned} \tag{C.6}$$

where ℓ divides k,

$$\delta_s(\ell) = \frac{1}{\ell} \log\left\{ s^2 \left[1 + \log\left(1 + \frac{\alpha^\ell \beta^\ell}{s^2} \right) \right] \right\} \tag{C.7}$$

and $\Delta_k'(\ell)$ is as defined in Subsection 2.4. But $\hat{H}_\ell(X_i^\ell, Y_i^\ell)/\ell$ can be decomposed also as

$$\frac{\hat{H}(X_i^\ell)}{\ell} + \frac{\hat{H}(Y_i^\ell|X_i^\ell)}{\ell} \ge \rho_{\rm LZ}(x_{ik+1}^{ik+k}) + \rho_{\rm LZ}(y_{ik+1}^{ik+k}|x_{ik+1}^{ik+k}) - 2\Delta_k'(\ell) \tag{C.8}$$

as well as

$$\frac{\hat{H}(Y_i^\ell)}{\ell} + \frac{\hat{H}(X_i^\ell|Y_i^\ell)}{\ell} \ge \rho_{\rm LZ}(y_{ik+1}^{ik+k}) + \rho_{\rm LZ}(x_{ik+1}^{ik+k}|y_{ik+1}^{ik+k}) - 2\Delta_k'(\ell), \tag{C.9}$$

which together imply that

$$\begin{aligned} \frac{L_i}{k} &\ge \max\{\rho_{\rm LZ}(x_{ik+1}^{ik+k}, y_{ik+1}^{ik+k}), \rho_{\rm LZ}(x_{ik+1}^{ik+k}) + \rho_{\rm LZ}(y_{ik+1}^{ik+k}|x_{ik+1}^{ik+k}), \\ &\qquad \rho_{\rm LZ}(y_{ik+1}^{ik+k}) + \rho_{\rm LZ}(x_{ik+1}^{ik+k}|y_{ik+1}^{ik+k})\} - \delta_s(\ell) - 2\Delta_k'(\ell) \\ &= \rho_{\rm LZ}^+(x_{ik+1}^{ik+k}, y_{ik+1}^{ik+k}) - \delta_s(\ell) - 2\Delta_k'(\ell). \end{aligned} \tag{C.10}$$

Thus, the s-state compressibility of (x^n, y^n) is lower bounded as follows:

$$\varrho_s(x^n, y^n) \geq \frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{\mathrm{LZ}}^+(x_{ik+1}^{ik+k}, y_{ik+1}^{ik+k}) - \delta_s(\ell) - 2\Delta_k'(\ell), \tag{C.11}$$

which leads to

$$\varrho_s(\boldsymbol{x}, \boldsymbol{y}) = \limsup_{n\to\infty} \varrho_s(x^n, y^n) \geq \limsup_{n\to\infty} \frac{k}{n} \sum_{i=0}^{n/k-1} \rho_{\mathrm{LZ}}^+(x_{ik+1}^{ik+k}, y_{ik+1}^{ik+k}) - \delta_s(\ell) - 2\Delta_k'(\ell). \tag{C.12}$$

Upon taking k (and then ℓ) to infinity (yet keeping s fixed), we obtain

$$\varrho_s(\boldsymbol{x}, \boldsymbol{y}) \geq \rho^+(\boldsymbol{x}, \boldsymbol{y}), \tag{C.13}$$

and so,

$$\varrho_\infty(\boldsymbol{x}, \boldsymbol{y}) = \lim_{s\to\infty} \varrho_s(\boldsymbol{x}, \boldsymbol{y}) \geq \rho^+(\boldsymbol{x}, \boldsymbol{y}), \tag{C.14}$$

which together with (C.5), yields

$$\rho^+(\boldsymbol{x}, \boldsymbol{y}) \leq \rho^-(\boldsymbol{x}, \boldsymbol{y}). \tag{C.15}$$

Consequently, $\rho^+(\boldsymbol{x}, \boldsymbol{y}) = \rho^-(\boldsymbol{x}, \boldsymbol{y}) = \rho(\boldsymbol{x}, \boldsymbol{y})$, and the proof of Theorem 2 is complete.

References

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

[2] S. C. Draper, "Universal incremental Slepian–Wolf coding," *Proc. Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, October 2004.

[3] 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.

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

[5] N. Merhav, "Universal decoding for source-channel coding with side information," *Communications in Information and Systems*, vol. 16, no. 1, pp. 17-58, 2016.

[6] N. Merhav, "A universal ensemble for sample-wise lossy compression," *Entropy*, 2023, 25(8), 1199; `https://doi.org/10.3390/e25081199`, August 2023.

[7] N. Shulman, *Communication over an Unknown Channel in Common Broadcasting*, Ph.D. thesis, Department of Electrical Engineering – Systems, Tel Aviv University, 2003.

[8] N. Shulman and M. Feder, "Source broadcasting with an unknown amount of receiver side information," *Proc. 2002 IEEE Information Theory Workshop*, Bangalore, India, pp. 127–130, October 2002.

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

[10] 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.

[11] J. Ziv, "Fixed-rate encoding of individual sequences with side information", *IEEE Transactions on Information Theory*, vol. IT–30, no. 2, pp. 348–452, March 1984.

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

[13] 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.

[14] 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.