

# Successive Refinement for Lossy Compression of Individual Sequences

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](mailto:merhav@ee.technion.ac.il)

## Abstract

We consider the problem of successive-refinement coding for lossy compression of individual sequences, namely, compression in two stages, where in the first stage, a coarse description at a relatively low rate is sent from the encoder to the decoder, and in the second stage, additional coding rate is allocated in order to refine the description and thereby improve the reproduction. Our main result is in establishing outer bounds (converse theorems) for the rate region where we limit the encoders to be finite-state machines in the spirit of Ziv and Lempel's 1978 model. The matching achievability scheme is conceptually straightforward. We also consider the more general multiple description coding problem on a similar footing and propose achievability schemes that are analogous to the well-known El Gamal-Cover and the Zhang-Berger achievability schemes of memoryless sources and additive distortion measures.

## 1 Introduction

The notion of *successive refinement* of information refers to systems in which the reconstruction of the source occurs in multiple stages. In these systems, a single encoder encodes the source and communicates with either one decoder or multiple decoders in a step-by-step manner. At each stage, the encoder transmits a portion of the source information to the corresponding decoder, which has access also to all previous transmissions. Each decoder uses all available transmissions to reconstruct the source, possibly incorporating additional side information. The quality of the reconstruction at each stage (or by each decoder) is evaluated based on a predefined distortion measure. One of the motivations of this hierarchical structure is to allow scalability that meets the available channel resources of the various users who receive the compressed information or

even to adapt to time-varying channel conditions of a single user. Several studies have addressed the successive refinement problem for probabilistic sources, most notably, memoryless sources, see, e.g., [1], where necessary and sufficient conditions for simultaneously achieving the rate-distortion function at all stages, [2] and [3], where the rate-distortion region was fully characterized in general, and [4], where source coding error exponents were derived. In some later works, such as [5], [6], [7], and [8], successive refinement coding was considered also with the incorporation of side information. Successive-refinement coding is also an important special case of the so called *multiple description coding*, which in its simplest form, consists of two encoders that send two different individual descriptions of the source to two separate respective decoders (that do not cooperate with each other), and the compressed bit-streams pertaining to those descriptions are also combined and sent to yet another decoder, whose role is to produce a better reconstruction than both of those of the two individual decoders. The problem of fully characterizing the rate-distortion region of the multiple description coding problem is open in its general form, and there are certain inner and outer bounds to this region, see Chapter 13 of [9] for some details and references therein.

In this paper, we focus on successive refinement of information in the context of individual sequences, namely, deterministic source sequences, as opposed to the traditional setting of random sequences governed by a certain probabilistic mechanism. In that sense, this work can be viewed as an additional step in the development of multiuser information theory for individual sequences, following a series of earlier works of this flavor, initiated by Ziv and Lempel in [10], [11], [12], and [13], and continued by others in many articles, such as [14], [15], [16], [17], [18], and [19]. In particular, we consider the problem of successive-refinement coding for lossy compression of individual sequences, in two stages, where in the first stage, a coarse description at a relatively low rate is sent from the encoder to the decoder, and in the second stage, additional coding rate is allocated in order to refine the description and thereby improve the reproduction. Our main results are in establishing outer bounds (converse theorems) for the rate region with individual sequences, where we limit the encoders to be finite-state machines similarly as in [11], [12], [13] and others. The compatible achievability scheme is conceptually straightforward, and so, we believe that the deeper and more interesting part of the contribution of this work is in the converse theorems, namely, in the outer bounds. Our results are formulated and proved for two stages of coding, but their extension to any fixed number of stages is straightforward. These results can also be viewed

as an extension of the fixed-distortion results of [14] to successive-refinement coding.

We also consider the more general multiple description coding problem and propose achievability schemes that are analogous to the well-known El Gamal-Cover [20] and the Zhang-Berger [21] achievability schemes of memoryless sources and additive distortion measures. There is a clear parallelism between the rate expressions that we obtain and those of [20] and [21], including those that are associated with the gaps between the outer bound and the corresponding inner bounds.

The outline of the remaining part of this paper is as follows. In Section 2, we establish notation conventions and formulate the problem setting. In Section 3, we provide some general background on the LZ algorithm as well as on its conditional form. Section 4 is devoted to the above mentioned successive refinement outer bound, and finally, in Section 5, we address the more general multiple description problem.

## 2 Notation Conventions and Problem Formulation

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 dimensions. More specifically, for a given positive integer,  $n$ , the source vector  $(x_1, x_2, \dots, x_n)$ , with components,  $x_i$ ,  $i = 1, 2, \dots, n$ , from a finite-alphabet,  $\mathcal{X}$ , will be denoted by  $x^n$ . The set of all such  $n$ -vectors will be denoted by  $\mathcal{X}^n$ , which is the  $n$ -th order Cartesian power of the single-letter source alphabet,  $\mathcal{X}$ . Likewise, reproduction vectors of length  $n$ , such as  $(\hat{x}_1, \dots, \hat{x}_n)$  and  $(\tilde{x}_1, \dots, \tilde{x}_n)$ , with components,  $\hat{x}_i$  and  $\tilde{x}_i$ ,  $i = 1, \dots, n$ , from finite-alphabets,  $\hat{\mathcal{X}}$  and  $\tilde{\mathcal{X}}$ , will be denoted by  $\hat{x}^n \in \hat{\mathcal{X}}^n$  and  $\tilde{x}^n \in \tilde{\mathcal{X}}^n$ , respectively. An infinite source sequence  $(x_1, x_2, \dots)$  will be denoted by  $\mathbf{x}$ . The cardinalities of  $\mathcal{X}$ ,  $\hat{\mathcal{X}}$ , and  $\tilde{\mathcal{X}}$  will be denoted by  $\alpha$ ,  $\beta$ , and  $\gamma$ , respectively. The value  $\alpha = \infty$  is allowed, to incorporate countably infinite and continuous source alphabets. By contrast,  $\beta$  and  $\gamma$  will always be finite.

For two positive integers  $i$  and  $j$ , where  $i \leq j$ , the notation  $x_i^j$  will be used to denote the substring  $(x_i, x_{i+1}, \dots, x_j)$ . For  $i = 1$ , the subscript ‘1’ will be omitted, and so, the shorthand notation of  $(x_1, x_2, \dots, x_n)$  will be  $x^n$  as mentioned before. Similar conventions will apply to other sequences.

Probability distributions will be denoted by the letter  $P$  with possible subscripts, depending on the context. The indicator function for an event  $\mathcal{A}$  will be denoted by  $\mathcal{I}\{\mathcal{A}\}$ , that is,  $\mathcal{I}\{\mathcal{A}\} = 1$  if  $\mathcal{A}$  occurs, and  $\mathcal{I}\{\mathcal{A}\} = 0$ , if not. The logarithmic function,  $\log x$ , will be understood to be defined to the base 2. Logarithms to the base  $e$  will be denoted by  $\ln$ . Let  $d_1 : \mathcal{X}^n \times \hat{\mathcal{X}}^n \rightarrow \mathbb{R}^+$  and  $d_2 : \mathcal{X}^n \times \tilde{\mathcal{X}}^n \rightarrow \mathbb{R}^+$  be two arbitrary distortion functions between source vectors,  $x^n$ , and corresponding reproduction vectors,  $\hat{x}^n$  and  $\tilde{x}^n$ , respectively.

The successive refinement encoder model is as follows. It is composed of a cascade of two encoders: a reproduction encoder (namely, a vector quantizer) followed by a lossless encoder. The input to the reproduction encoder is the source sequence  $x^n$  and the output is a pair of reproduction vectors,  $(\hat{x}^n, \tilde{x}^n) \in \hat{\mathcal{X}}^n \times \tilde{\mathcal{X}}^n$  that obeys the distortion constraints,  $d_1(x^n, \hat{x}^n) \leq nD_1$  and  $d_2(x^n, \tilde{x}^n) \leq nD_2$ , where  $D_1 \geq 0$  and  $D_2 \geq 0$  are prescribed normalized distortion levels. There are no particular restrictions imposed on the distortion functions. We denote

$$\mathcal{B}(x^n) = \{(\hat{x}^n, \tilde{x}^n) : d_1(x^n, \hat{x}^n) \leq nD_1, d_2(x^n, \tilde{x}^n) \leq nD_2\}. \quad (1)$$

The mapping  $\mathcal{X}^n \rightarrow \hat{\mathcal{X}}^n \times \tilde{\mathcal{X}}^n$ , employed by the reproduction encoder, is arbitrary and not limited. The pair  $(\hat{x}^n, \tilde{x}^n)$  serves as an input to the lossless encoder.

As for the lossless encoder, we follow the same modeling approach as in [11], but with a few adjustments to make it suitable to successive refinement. In particular, the lossless encoder is defined by a set

$$E = (\hat{\mathcal{X}}, \tilde{\mathcal{X}}, \mathcal{U}, \mathcal{V}, \mathcal{S}, \mathcal{Z}, f_1, f_2, g_1, g_2),$$

where  $\hat{\mathcal{X}}$  and  $\tilde{\mathcal{X}}$  are as before;  $\mathcal{U}$  and  $\mathcal{V}$  are two sets of variable-length binary strings, which both include the empty string  $\lambda$  of length zero;  $\mathcal{S}$  and  $\mathcal{Z}$  are two sets of states, each one containing  $q$  states;  $f_1 : \mathcal{S} \times \hat{\mathcal{X}} \rightarrow \mathcal{U}$  and  $f_2 : \mathcal{Z} \times \hat{\mathcal{X}} \times \tilde{\mathcal{X}} \rightarrow \mathcal{V}$  are the encoder output functions, and finally,  $g_1 : \mathcal{S} \times \hat{\mathcal{X}} \rightarrow \mathcal{S}$  and  $g_2 : \mathcal{Z} \times \hat{\mathcal{X}} \times \tilde{\mathcal{X}} \rightarrow \mathcal{Z}$  are two next-state functions. When the lossless encoder is fed by a sequence of pairs  $(\hat{x}_1, \tilde{x}_1), (\hat{x}_2, \tilde{x}_2), \dots$ , it outputs a corresponding sequence of pairs of binary strings,  $(u_1, v_1), (u_2, v_2), \dots$  according to the following recursive mechanism. For  $t = 1, 2, \dots$ :

$$u_t = f_1(s_t, \hat{x}_t) \quad (2)$$

$$s_{t+1} = g_1(s_t, \hat{x}_t) \quad (3)$$

$$v_t = f_2(z_t, \hat{x}_t, \tilde{x}_t) \quad (4)$$

$$z_{t+1} = g_2(z_t, \hat{x}_t, \tilde{x}_t), \quad (5)$$

where the initial states,  $s_1$  and  $z_1$ , are assumed arbitrary fixed members of  $\mathcal{S}$  and  $\mathcal{Z}$ , respectively. Similarly as in [11], we adopt the extended notation, of  $f_1(s_1, \hat{x}^n)$  for  $u^n$ ,  $g_1(s_1, \hat{x}^n)$  for  $s_{n+1}$ , and similar notations associated with  $f_2$  and  $g_2$ .

An encoder  $E$  is said to be information lossless if for every positive integer  $k$ , the vector  $(s_1, f_1(s_1, \hat{x}^k), g_1(s_1, \hat{x}^k))$  uniquely determines  $\hat{x}^k$  and likewise,

$$(s_1, z_1, f_1(s_1, \hat{x}^k), f_2(z_1, \hat{x}^k, \tilde{x}^k), g_1(s_1, \tilde{x}^k), g_2(z_1, \hat{x}^k, \tilde{x}^k))$$

uniquely determines  $(\hat{x}^k, \tilde{x}^k)$ . Let  $\mathcal{E}(q)$  be the set of all information lossless encoders,  $\{E\}$ , with  $|\mathcal{S}| \leq q$  and  $|\mathcal{Z}| \leq q$ .

Given a lossless encoder  $E$  and a pair of inputs  $(\hat{x}^n, \tilde{x}^n)$ , we define

$$[\rho_E(\hat{x}^n)]_1 = \frac{L(u^n)}{n} = \frac{L[f_1(s_1, \hat{x}^n)]}{n} \quad (6)$$

and

$$[\rho_E(\hat{x}^n, \tilde{x}^n)]_2 = \frac{L(u^n) + L(v^n)}{n} = \frac{L[f_1(s_1, \hat{x}^n)] + L[f_2(z_1, \hat{x}^n, \tilde{x}^n)]}{n}, \quad (7)$$

where  $L(u^n) = \sum_{i=1}^n l(u_i)$ ,  $l(u_i)$  being the length (in bits) of the binary string  $u_i$ , and similarly for  $L(v^n)$ . Recall that for the empty string,  $\lambda$ , we define  $l(\lambda) = 0$ . The achievable rate region for  $x^n$  that is associated with  $E$  is defined as

$$\mathcal{R}_E(x^n) = \bigcup_{(\hat{x}^n, \tilde{x}^n) \in \mathcal{B}(x^n)} \{(R_1, R_2) : R_1 \geq [\rho_E(\hat{x}^n)]_1, R_1 + R_2 \geq [\rho_E(\hat{x}^n, \tilde{x}^n)]_2\}, \quad (8)$$

and the  $q$ -state achievable rate region for  $x^n$  is defined as

$$\mathcal{R}_q(x^n) = \bigcup_{E \in \mathcal{E}(q)} \mathcal{R}_E(x^n). \quad (9)$$

The rationale behind the union operations in these definitions is that they are two-dimensional set-theoretic analogues of the minimization operations over the appropriate encoders  $\{E\}$  and reproduction vectors,  $\{\hat{x}^n\}$ , that appear for a single coding rate, like in [11] and [14], as can be noted from the simple relationship:

$$\begin{aligned} & \left\{ R : R \geq \min_{E \in \{\text{all } q\text{-state encoders}\}} \min_{\{\hat{x}^n : d(x^n, \hat{x}^n) \leq nD\}} \rho_E(\hat{x}^n) \right\} \\ = & \bigcup_{E \in \{\text{all } q\text{-state encoders}\}} \bigcup_{\{\hat{x}^n : d(x^n, \hat{x}^n) \leq nD\}} \{R : R \geq \rho_E(\hat{x}^n)\} \end{aligned} \quad (10)$$

for a given generic distortion function  $d$  and distortion level  $D$ .

For later use, we also define the joint empirical distribution of  $\ell$ -blocks of  $(\hat{x}_{i\ell+1}^{i\ell+\ell}, \tilde{x}_{i\ell+1}^{i\ell+\ell})$ ,  $\ell = 0, 1, \dots, n/\ell - 1$ , provided that  $\ell$  divides  $n$ . Specifically, consider the empirical distribution,  $\hat{P} = \{\hat{P}(\hat{x}^\ell, \tilde{x}^\ell), \hat{x}^\ell \in \hat{\mathcal{X}}^\ell, \tilde{x}^\ell \in \tilde{\mathcal{X}}^\ell\}$ , of pairs of  $\ell$ -vectors, defined as

$$\hat{P}(\hat{x}^\ell, \tilde{x}^\ell) = \frac{\ell}{n} \sum_{i=0}^{n/\ell-1} \mathcal{I}\{\hat{x}_{i\ell+1}^{i\ell+\ell} = \hat{x}^\ell, \tilde{x}_{i\ell+1}^{i\ell+\ell} = \tilde{x}^\ell\}, \quad \hat{x}^\ell \in \hat{\mathcal{X}}^\ell, \tilde{x}^\ell \in \tilde{\mathcal{X}}^\ell \quad (11)$$

Let  $H(\hat{X}^\ell, \tilde{X}^\ell)$  denote the joint empirical entropy of an auxiliary pair of random  $\ell$ -vectors,  $(\hat{X}^\ell, \tilde{X}^\ell)$ , induced by  $\hat{P}$ , that is,

$$H(\hat{X}^\ell, \tilde{X}^\ell) = - \sum_{(\hat{x}^\ell, \tilde{x}^\ell) \in \hat{\mathcal{X}}^\ell \times \tilde{\mathcal{X}}^\ell} \hat{P}(\hat{x}^\ell, \tilde{x}^\ell) \log \hat{P}(\hat{x}^\ell, \tilde{x}^\ell). \quad (12)$$

Accordingly,  $H(\hat{X}^\ell)$  and  $H(\tilde{X}^\ell | \hat{X}^\ell)$  will denote the corresponding marginal empirical entropy of  $\hat{X}^\ell$  and the conditional empirical entropy of  $\tilde{X}^\ell$  given  $\hat{X}^\ell$ .

Our objective is to provide inner and outer bounds to the achievable rate region and show that they asymptotically coincide in the limit of large  $n$  followed by a limit of large  $q$ , in analogy to the asymptotic regime of [11].

### 3 Background

Before the exposition of the main results and their proofs, we revisit key terms and details related to the 1978 version of the LZ algorithm, also known as the LZ78 algorithm [11], which is the

central building block in this work. The incremental parsing procedure of the LZ78 algorithm is a sequential parsing process applied to a finite-alphabet input vector,  $\hat{x}^n$ . According to 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  $\hat{x}^{15} = \text{abbabaabbbaaaba}$  results in a,b,ba,baa,bb,aa,ab,aa. Let  $c(\hat{x}^n)$  denote the number of phrases in  $\hat{x}^n$  resulting from the incremental parsing procedure (in the above example,  $c(\hat{x}^{15}) = 8$ ). Furthermore, let  $LZ(\hat{x}^n)$  denote the length of the LZ78 binary compressed code for  $\hat{x}^n$ . According to [11, Theorem 2], the following inequality holds:

$$\begin{aligned}
LZ(\hat{x}^n) &\leq [c(\hat{x}^n) + 1] \log\{2\beta[c(\hat{x}^n) + 1]\} \\
&= c(\hat{x}^n) \log[c(\hat{x}^n) + 1] + c(\hat{x}^n) \log(2\beta) + \log\{2\beta[c(\hat{x}^n) + 1]\} \\
&= c(\hat{x}^n) \log c(\hat{x}^n) + c(\hat{x}^n) \log \left[1 + \frac{1}{c(\hat{x}^n)}\right] + c(\hat{x}^n) \log(2\beta) + \log\{2\beta[c(\hat{x}^n) + 1]\} \\
&\leq c(\hat{x}^n) \log c(\hat{x}^n) + \log e + \frac{n(\log \beta) \log(2\beta)}{(1 - \varepsilon_n) \log n} + \log[2\beta(n + 1)] \\
&\triangleq c(\hat{x}^n) \log c(\hat{x}^n) + n \cdot \epsilon(n),
\end{aligned} \tag{13}$$

where we remind that  $\beta$  is the cardinality of  $\hat{\mathcal{X}}$ , and where both  $\varepsilon_n$  and  $\epsilon(n)$  tends to zero as  $n \rightarrow \infty$ . In other words, the LZ code-length for  $\hat{x}^n$  is upper bounded by an expression whose main term is  $c(\hat{x}^n) \log c(\hat{x}^n)$ . On the other hand,  $c(\hat{x}^n) \log c(\hat{x}^n)$  is also known to be the main term of a lower bound (see Theorem 1 of [11]) to the shortest code-length attainable by any information lossless finite-state encoder with no more than  $q$  states, provided that  $\log(q^2)$  is very small compared to  $\log c(\hat{x}^n)$ . In view of these facts, we henceforth refer to  $c(\hat{x}^n) \log c(\hat{x}^n)$  as the unnormalized *LZ complexity* of  $\hat{x}^n$  whereas the normalized LZ complexity is defined as

$$\rho_{\text{LZ}}(\hat{x}^n) \triangleq \frac{c(\hat{x}^n) \log c(\hat{x}^n)}{n}. \tag{14}$$

A useful inequality, that relates the empirical entropy of non-overlapping  $\ell$ -blocks of  $\hat{x}^n$  (where  $\ell$  divides  $n$ ) and  $\rho_{\text{LZ}}(\hat{x}^n)$  (see, for example, eq. (26) of [22]), is the following:

$$\frac{H(\hat{X}^\ell)}{\ell} \geq \rho_{\text{LZ}}(\hat{x}^n) - \frac{\log[4\beta^{2\ell}] \log \beta}{(1 - \varepsilon_n) \log n} - \frac{\beta^{2\ell} \log[4\beta^{2\ell}]}{n} - \frac{1}{\ell}$$

$$\triangleq \rho_{\text{LZ}}(x^n) - \delta_n(\ell), \quad (15)$$

It is obtained from the fact that the Shannon code for  $\ell$ -blocks can be implemented using a finite-state encoder with no more than  $\beta^\ell$  states. Specifically, for a block code of length  $\ell$  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  $\beta$  is then  $\sum_{i=0}^{\ell-1} \beta^i = (\beta^\ell - 1)/(\beta - 1) < \beta^\ell$ . Therefore, the code-length of this Shannon code must comply with the lower bound of Theorem 1 in [11]. Note that  $\lim_{n \rightarrow \infty} \delta_n(\ell) = 1/\ell$  and so,  $\lim_{\ell \rightarrow \infty} \lim_{n \rightarrow \infty} \delta_n(\ell) = 0$ . Clearly, it is possible to let  $\ell = \ell(n)$  increase with  $n$  slowly enough such that  $\delta_n(\ell(n)) \rightarrow 0$  as  $n \rightarrow \infty$ , in particular,  $\ell(n)$  should be  $o(\log n)$  for that purpose.

In [23], 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  $\hat{x}^n$  and  $\tilde{x}^n$ , let us apply the incremental parsing procedure of the LZ algorithm to the sequence of pairs  $((\hat{x}_1, \tilde{x}_1), (\hat{x}_2, \tilde{x}_2), \dots, (\hat{x}_n, \tilde{x}_n))$ . As mentioned before, according to this procedure, all phrases are distinct with a possible exception of the last phrase, which might be incomplete. Let  $c(\hat{x}^n, \tilde{x}^n)$  denote the number of distinct phrases. As an example (taken from [23]), let  $n = 6$  and consider the sequence pair  $(\hat{x}^6, \tilde{x}^6)$  along with its joint incremental parsing as follows:

$$\begin{aligned} \hat{x}^6 &= 0 \mid 1 \mid 0 \mid 1 \mid 0 \mid 1 \mid \\ \tilde{x}^6 &= 0 \mid 1 \mid 0 \mid 0 \mid 0 \mid 1 \mid \end{aligned} \quad (16)$$

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

$$\rho_{LZ}(\tilde{x}^n|\hat{x}^n) \triangleq \frac{1}{n} \sum_{l=1}^{c'(\hat{x}^n)} c_l(\tilde{x}^n|\hat{x}^n) \log c_l(\tilde{x}^n|\hat{x}^n). \quad (17)$$

In [23] it was shown that  $\rho_{LZ}(\tilde{x}^n|\hat{x}^n)$  is the main term of the compression ratio achieved by the conditional version of the LZ algorithm for compressing  $\tilde{x}^n$  in the presence of the side information  $\hat{x}^n$ , available to both encoder and decoder – see the compression scheme described in [23] (see also [15]), i.e., the length function,  $LZ(\tilde{x}^n|\hat{x}^n)$ , of the coding scheme proposed therein is upper bounded (in parallel to (13)) by

$$LZ(\tilde{x}^n|\hat{x}^n) \leq n\rho_{LZ}(\tilde{x}^n|\hat{x}^n) + n\hat{\epsilon}(n), \quad (18)$$

where  $\hat{\epsilon}(n) = O\left(\frac{\log(\log n)}{\log n}\right)$  (see eqs. (10) and (11) in [15]). On the other hand, analogously to [11, Theorem 1], it was shown in [16], that  $\rho_{LZ}(\tilde{x}^n|\hat{x}^n)$  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 (15) also extends to the conditional case as follows (see [16]):

$$\frac{H(\tilde{X}^\ell|\hat{X}^\ell)}{\ell} \geq \rho_{LZ}(\tilde{x}^n|\hat{x}^n) - \delta'_n(\ell), \quad (19)$$

where  $\delta'_n(\ell)$  is the same as  $\delta_n(\ell)$  except that  $\beta^\ell$  therein is replaced by  $(\beta\gamma)^\ell$  to accommodate the number of states associated with the conditional version of the aforementioned Shannon code applied to  $\ell$ -blocks. By the same token, we also have

$$\frac{H(\hat{X}^\ell, \tilde{X}^\ell)}{\ell} \geq \rho_{LZ}(\hat{x}^n, \tilde{x}^n) - \delta'_n(\ell). \quad (20)$$

## 4 The Outer Bound for Successive Refinement

Our main result for finite-state encoders is the following.

*Theorem 1.* For every  $x^n \in \mathcal{X}^n$ ,

$$\begin{aligned}
\mathcal{R}_q(x^n) &\subseteq \mathcal{R}_o(x^n) \\
&\triangleq \bigcup_{(\hat{x}^n, \tilde{x}^n) \in \mathcal{B}(x^n)} \mathcal{R}_{\text{LZ}}(\hat{x}^n, \tilde{x}^n),
\end{aligned} \tag{21}$$

where

$$\begin{aligned}
\mathcal{R}_{\text{LZ}}(\hat{x}^n, \tilde{x}^n) &\triangleq \left\{ (R_1, R_2) : \right. \\
R_1 &\geq \rho_{\text{LZ}}(\hat{x}^n) - \Delta_1(q, n), \\
R_1 + R_2 &\geq \rho_{\text{LZ}}(\hat{x}^n) + \rho_{\text{LZ}}(\tilde{x}^n | \hat{x}^n) - \Delta_2(q, n) \left. \right\},
\end{aligned} \tag{22}$$

where  $\Delta_1(q, n)$  and  $\Delta_2(q, n)$  are defined as

$$\Delta_1(q, n) = \frac{\log(4q^2) \log \beta}{(1 - \epsilon_n) \log n} + \frac{q^2 \log(4q^2)}{n} \tag{23}$$

with  $\epsilon_n \rightarrow 0$  as  $n \rightarrow \infty$ , and

$$\Delta_2(n, q) = \min_{\{\ell: \ell \text{ divides } n\}} \left\{ \delta_n(\ell) + \delta'_n(\ell) + \frac{1}{\ell} \log \left[ q^4 \left( 1 + \log \left[ 1 + \frac{\beta^\ell \gamma^\ell}{q^4} \right] \right) \right] \right\}. \tag{24}$$

**Discussion.** Several comments are now in order.

1. Since both  $\Delta_1(q, n)$  and  $\Delta_2(q, n)$  tend to zero as  $n \rightarrow \infty$  for fixed  $q$ , the asymptotic achievability of  $\mathcal{R}_o(x^n)$  is conceptually straightforward: Given an internal point,  $(R_1, R_2) \in \mathcal{R}_o(x^n)$ , there must be at least one pair  $(\hat{x}^n, \tilde{x}^n) \in \mathcal{B}(x^n)$  such that  $(R_1, R_2) \in \mathcal{R}_{\text{LZ}}(\hat{x}^n, \tilde{x}^n)$ . Upon finding such a pair, proceed as follows: At the first stage, apply LZ78 compression to  $\hat{x}^n$  at a coding rate of  $R_1 = \frac{LZ(\hat{x}^n)}{n}$  which is only slightly above  $\rho_{\text{LZ}}(\hat{x}^n)$  for large  $n$ , as discussed in Section 3. At the second stage, apply conditional LZ compression of  $\tilde{x}^n$  given  $\hat{x}^n$  as side information at both ends, at an incremental coding rate of  $R_2 = \frac{LZ(\tilde{x}^n | \hat{x}^n)}{n}$  which is close to  $\rho_{\text{LZ}}(\tilde{x}^n | \hat{x}^n)$ , as also explained in Section 3, and then the total rate,  $R_1 + R_2$ , is about  $\rho_{\text{LZ}}(\hat{x}^n) + \rho_{\text{LZ}}(\tilde{x}^n | \hat{x}^n)$ . Similarly as in [11], there is still a certain gap between the achievability and the converse theorem, because the achievability requires

encoders whose number of states is not small compared to  $n$ , whereas the converse is significant when  $q$  is very small relative to  $n$ . As in [11], this gap can be closed in the asymptotic limit of large  $q$  by partitioning the sequence into non-overlapping blocks and starting over the LZ compression mechanism in each block separately. We will address this point in detail later on.

2. Considering that the achievability is conceptually straightforward, as explained in item no. 1 above, the interesting and deeper result is the converse theorem. Since the second stage encoder receives both  $\hat{x}^n$  and  $\tilde{x}^n$  as inputs, it is immediate to lower bound the total coding rate, at the second stage, in terms of the joint compressibility of  $(\hat{x}^n, \tilde{x}^n)$ , namely by  $\rho_{\text{LZ}}(\hat{x}^n, \tilde{x}^n)$ , but recall that the first-stage encoder must have already allocated a rate at least as large as  $\rho_{\text{LZ}}(\hat{x}^n)$ , then in order to meet a lower bound of  $\rho_{\text{LZ}}(\hat{x}^n, \tilde{x}^n)$  on the total coding rate, the incremental rate,  $R_2$ , of the second stage must not exceed  $\rho_{\text{LZ}}(\hat{x}^n, \tilde{x}^n) - \rho_{\text{LZ}}(\hat{x}^n)$ , and there is no apparent way to achieve such a coding rate, as far as the author can see. Nonetheless, since we can also lower bound the total rate of both stages by  $\rho_{\text{LZ}}(\hat{x}^n) + \rho_{\text{LZ}}(\tilde{x}^n|\hat{x}^n)$ , then the achievability becomes obvious, as said. This point is not trivial because there is no chain rule that applies to the LZ complexities of arbitrary finite sequences. The proof that  $\rho_{\text{LZ}}(\hat{x}^n) + \rho_{\text{LZ}}(\tilde{x}^n|\hat{x}^n)$  also serves as a lower bound (essentially), requires a certain manipulation by using a generalized Kraft inequality and passing via empirical entropies, as can be seen in the proof.

3. The choice of  $\hat{x}^n$  exhibits a trade-off between the coding rate of the first stage and the incremental rate at the second stage because  $\hat{x}^n$  is both compressed at the first stage and serves as side information at the second stage, so there might be a certain tension between selecting  $\hat{x}^n$  for having small  $\rho_{\text{LZ}}(\hat{x}^n)$  and selecting it for small  $\rho_{\text{LZ}}(\tilde{x}^n|\hat{x}^n)$ . Of course, an analogous tension exists also in successive refinement for memoryless sources [3]. The reproduction encoder must select  $(\hat{x}^n, \tilde{x}^n) \in \mathcal{B}(x^n)$  that best compromises these criteria.

4. The results extend straightforwardly to any finite number of stages, where at each stage one applies conditional LZ compression of the current reproduction given all previous reproductions.

*Proof of Theorem 1.* We begin from the first stage. By definition, if  $(R_1, R_2) \in \mathcal{R}_q(x^n)$ , then there must exist an encoder  $E \in \mathcal{E}(q)$  and  $(\hat{x}^n, \tilde{x}^n) \in \mathcal{B}(x^n)$  such that  $R_1 \geq [\rho_E(\hat{x}^n)]_1$  and  $R_1 + R_2 \geq [\rho_E(\hat{x}^n, \tilde{x}^n)]_2$ . Now, according to Theorem 1 of [11]:

$$\begin{aligned}
[\rho_E(\hat{x}^n)]_1 &\geq \frac{c(\hat{x}^n) + q^2}{n} \cdot \log \left[ \frac{c(\hat{x}^n) + q^2}{4q^2} \right] + \frac{2q^2}{n} \\
&> \frac{c(\hat{x}^n) + q^2}{n} \cdot \log[c(\hat{x}^n) + q^2] - \frac{c(\hat{x}^n) + q^2}{n} \log(4q^2) \\
&> \frac{c(\hat{x}^n) \log c(\hat{x}^n)}{n} - \frac{c(\hat{x}^n) \log(4q^2)}{n} - \frac{q^2 \log(4q^2)}{n} \\
&\geq \frac{c(\hat{x}^n) \log c(\hat{x}^n)}{n} - \frac{\log(4q^2) \log \beta}{(1 - \epsilon_n) \log n} - \frac{q^2 \log(4q^2)}{n} \\
&= \rho_{\text{LZ}}(\hat{x}^n) - \Delta_1(q, n),
\end{aligned} \tag{25}$$

where  $\lim_{n \rightarrow \infty} \epsilon_n = 0$  and the last inequality is an application of eq. (6) in [11]. Since  $R_1 \geq [\rho_E(\hat{x}^n)]_1$ , it follows that

$$R_1 \geq \rho_{\text{LZ}}(\hat{x}^n) - \Delta_1(q, n). \tag{26}$$

Moving on to the combined encoder of both stages, consider the following. According to Lemma 2 of [11] and due to the postulated information losslessness, the combined encoder, which has  $q^2$  states, must obey the following generalized Kraft inequality:

$$\sum_{(\hat{x}^\ell, \tilde{x}^\ell) \in \hat{\mathcal{X}}^\ell \times \tilde{\mathcal{X}}^\ell} \exp_2 \left\{ -[\min_{s \in \mathcal{S}} L[f_1(s, \hat{x}^\ell)] + \min_{z \in \mathcal{Z}} L[f_2(z, \hat{x}^\ell, \tilde{x}^\ell)]] \right\} \leq q^4 \left( 1 + \log \left[ 1 + \frac{\beta^\ell \gamma^\ell}{q^4} \right] \right). \tag{27}$$

This implies that the description length at the output of this encoder is lower bounded as follows.

$$\begin{aligned}
n(R_1 + R_2) &\geq n[\rho_E(\hat{x}^n, \tilde{x}^n)]_2 \\
&= L(u^n) + L(v^n) \\
&= \sum_{t=1}^n \{L[f_1(s_t, \hat{x}_t)] + L[f_2(z_t, \hat{x}_t, \tilde{x}_t)]\} \\
&= \sum_{m=0}^{n/\ell-1} \sum_{j=1}^{\ell} \{L[f_1(s_{m\ell+j}, \hat{x}_{m\ell+j})] + L[f_2(z_{m\ell+j}, \hat{x}_{m\ell+j}, \tilde{x}_{m\ell+j})]\}
\end{aligned}$$

$$\begin{aligned}
&= \sum_{m=0}^{n/\ell-1} \{L[f_1(s_{m\ell+1}, \hat{x}_{m\ell+1}^{m\ell+\ell})] + L[f_2(z_{m\ell+1}, \hat{x}_{m\ell+1}^{m\ell+\ell}, \tilde{x}_{m\ell+1}^{m\ell+\ell})]\} \\
&\geq \sum_{m=0}^{n/\ell-1} \left\{ \min_{s \in \mathcal{S}} L[f_1(s, \hat{x}_{m\ell+1}^{m\ell+\ell})] + \min_{z \in \mathcal{Z}} L[f_2(z, \hat{x}_{m\ell+1}^{m\ell+\ell}, \tilde{x}_{m\ell+1}^{m\ell+\ell})] \right\} \\
&= \frac{n}{\ell} \sum_{(\hat{x}^\ell, \tilde{x}^\ell) \in \hat{\mathcal{X}}^\ell \times \tilde{\mathcal{X}}^\ell} \hat{P}(\hat{x}^\ell, \tilde{x}^\ell) \cdot \left\{ \min_{s \in \mathcal{S}} L[f_1(s, \hat{x}^\ell)] + \min_{z \in \mathcal{Z}} L[f_2(z, \hat{x}^\ell, \tilde{x}^\ell)] \right\} \quad (28)
\end{aligned}$$

and so,

$$R_1 + R_2 \geq \frac{1}{\ell} \sum_{(\hat{x}^\ell, \tilde{x}^\ell) \in \hat{\mathcal{X}}^\ell \times \tilde{\mathcal{X}}^\ell} \hat{P}(\hat{x}^\ell, \tilde{x}^\ell) \cdot \left\{ \min_{s \in \mathcal{S}} L[f_1(s, \hat{x}^\ell)] + \min_{z \in \mathcal{Z}} L[f_2(z, \hat{x}^\ell, \tilde{x}^\ell)] \right\}. \quad (29)$$

Now, by the generalized Kraft inequality above,

$$\begin{aligned}
&q^4 \left( 1 + \log \left[ 1 + \frac{\beta^\ell \gamma^\ell}{q^4} \right] \right) \\
&\geq \sum_{(\hat{x}^\ell, \tilde{x}^\ell) \in \hat{\mathcal{X}}^\ell \times \tilde{\mathcal{X}}^\ell} \exp_2 \left\{ - \left( \min_{s \in \mathcal{S}} L[f_1(s, \hat{x}^\ell)] + \min_{z \in \mathcal{Z}} L[f_2(z, \hat{x}^\ell, \tilde{x}^\ell)] \right) \right\} \\
&= \sum_{(\hat{x}^\ell, \tilde{x}^\ell) \in \hat{\mathcal{X}}^\ell \times \tilde{\mathcal{X}}^\ell} \hat{P}(\hat{x}^\ell, \tilde{x}^\ell) \cdot \exp_2 \left\{ - \left( \min_{s \in \mathcal{S}} L[f_1(s, \hat{x}^\ell)] + \min_{z \in \mathcal{Z}} L[f_2(z, \hat{x}^\ell, \tilde{x}^\ell)] \right) - \log \hat{P}(\hat{x}^\ell, \tilde{x}^\ell) \right\} \\
&\geq \exp_2 \left\{ - \sum_{(\hat{x}^\ell, \tilde{x}^\ell) \in \hat{\mathcal{X}}^\ell \times \tilde{\mathcal{X}}^\ell} \hat{P}(\hat{x}^\ell, \tilde{x}^\ell) \cdot \left( \min_{s \in \mathcal{S}} L[f_1(s, \hat{x}^\ell)] + \min_{z \in \mathcal{Z}} L[f_2(z, \hat{x}^\ell, \tilde{x}^\ell)] \right) + H(\hat{X}^\ell, \tilde{X}^\ell) \right\},
\end{aligned}$$

where the last inequality follows from the convexity of the exponential function and Jensen's inequality. This yields

$$\begin{aligned}
&\log \left\{ q^4 \left( 1 + \log \left[ 1 + \frac{\beta^\ell \gamma^\ell}{q^4} \right] \right) \right\} \\
&\geq H(\hat{X}^\ell, \tilde{X}^\ell) - \sum_{(\hat{x}^\ell, \tilde{x}^\ell) \in \hat{\mathcal{X}}^\ell \times \tilde{\mathcal{X}}^\ell} \hat{P}(\hat{x}^\ell, \tilde{x}^\ell) \cdot \left\{ \min_{s \in \mathcal{S}} L[f_1(s, \hat{x}^\ell)] + \min_{z \in \mathcal{Z}} L[f_2(z, \hat{x}^\ell, \tilde{x}^\ell)] \right\}, \quad (30)
\end{aligned}$$

implying that

$$R_1 + R_2 \geq \frac{L(u^n) + L(v^n)}{n}$$

$$\begin{aligned}
&\geq \frac{1}{\ell} \sum_{(\hat{x}^\ell, \tilde{x}^\ell) \in \hat{\mathcal{X}}^\ell \times \tilde{\mathcal{X}}^\ell} \hat{P}(\hat{x}^\ell, \tilde{x}^\ell) \cdot \left\{ \min_{s \in \mathcal{S}} L[f_1(s, \hat{x}^\ell)] + \min_{z \in \mathcal{Z}} L[f_2(z, \hat{x}^\ell, \tilde{x}^\ell)] \right\} \\
&\geq \frac{H(\hat{X}^\ell, \tilde{X}^\ell)}{\ell} - \frac{1}{\ell} \log \left\{ q^4 \left( 1 + \log \left[ 1 + \frac{\beta^\ell \gamma^\ell}{q^4} \right] \right) \right\} \\
&= \frac{H(\hat{X}^\ell)}{\ell} + \frac{H(\tilde{X}^\ell | \hat{X}^\ell)}{\ell} - \frac{1}{\ell} \log \left\{ q^4 \left( 1 + \log \left[ 1 + \frac{\beta^\ell \gamma^\ell}{q^4} \right] \right) \right\}
\end{aligned} \tag{31}$$

Now, according to eq. (15),

$$\frac{H(\hat{X}^\ell)}{\ell} \geq \rho_{\text{LZ}}(\hat{x}^n) - \delta_n(\ell). \tag{32}$$

Similarly, according to eq. (19),

$$\frac{H(\tilde{X}^\ell | \hat{X}^\ell)}{\ell} \geq \rho_{\text{LZ}}(\tilde{x}^n | \hat{x}^n) - \delta'_n(\ell), \tag{33}$$

and so,

$$R_1 + R_2 \geq \rho_{\text{LZ}}(\hat{x}^n) + \rho_{\text{LZ}}(\tilde{x}^n | \hat{x}^n) - \delta_n(\ell) - \delta'_n(\ell) - \frac{1}{\ell} \log \left[ q^4 \left( 1 + \log \left[ 1 + \frac{\beta^\ell \gamma^\ell}{q^4} \right] \right) \right]. \tag{34}$$

Maximizing this lower bound w.r.t.  $\ell$  yields

$$R_1 + R_2 \geq \rho_{\text{LZ}}(\hat{x}^n) + \rho_{\text{LZ}}(\tilde{x}^n | \hat{x}^n) - \Delta_2(n, q), \tag{35}$$

where

$$\Delta_2(n, q) = \min_{\{\ell: \ell \text{ divides } n\}} \left\{ \delta_n(\ell) + \delta'_n(\ell) + \frac{1}{\ell} \log \left[ q^4 \left( 1 + \log \left[ 1 + \frac{\beta^\ell \gamma^\ell}{q^4} \right] \right) \right] \right\}. \tag{36}$$

This completes the proof of Theorem 1.

Referring to the last part of comment no. 1 in the discussion that follows Theorem 1, we now address the gap in terms of the number of states. For an infinite source sequence  $\mathbf{x} = (x_1, x_2, \dots)$ , we define the  $q$ -state achievable rate region for  $\mathbf{x}$  as

$$\mathcal{R}_q(\mathbf{x}) = \bigcup_{m \geq 1} \bigcap_{n \geq m} \mathcal{R}_q(x^n), \quad (37)$$

and finally, the finite-state achievable rate region for  $\mathbf{x}$  is defined as

$$\mathcal{R}_\infty(\mathbf{x}) = \bigcup_{q \geq 1} \mathcal{R}_q(\mathbf{x}). \quad (38)$$

These definitions are two-dimensional counterparts of eqs. (2)–(4) in [11], where the finite-state (lossless) compressibility of  $\mathbf{x}$  is defined in several steps. In particular, the union over intersections in the definition of  $\mathcal{R}_q(\mathbf{x})$  is the set-theoretic analogue of the limit superior operation, and the union operation in the definition of  $\mathcal{R}_\infty(\mathbf{x})$  is parallel to a limit of  $q \rightarrow \infty$ .

Let  $k$  be a positive integer that divides  $n$  and consider the partition of  $\hat{x}^n$  and  $\tilde{x}^n$  into  $n/k$  blocks of length  $k$ , i.e.,  $\hat{x}_{kt+1}^{kt+k} = (\hat{x}_{kt+1}, \hat{x}_{kt+2}, \dots, \hat{x}_{kt+k})$  and  $\tilde{x}_{kt+1}^{kt+k} = (\tilde{x}_{kt+1}, \tilde{x}_{kt+2}, \dots, \tilde{x}_{kt+k})$ ,  $t = 0, 1, \dots, n/k - 1$ . Next, define:

$$\begin{aligned} \mathcal{R}_-^k(\hat{x}^n, \tilde{x}^n) = & \left\{ (R_1, R_2) : R_1 \geq \frac{k}{n} \sum_{t=0}^{n/k-1} \rho_{\text{LZ}}(\hat{x}_{kt+1}^{kt+k}) - \Delta_1(q, k), \right. \\ & \left. R_1 + R_2 \geq \frac{k}{n} \sum_{t=0}^{n/k-1} [\rho_{\text{LZ}}(\hat{x}_{kt+1}^{kt+k}) + \rho_{\text{LZ}}(\tilde{x}_{kt+1}^{kt+k} | \hat{x}_{kt+1}^{kt+k})] - \Delta_2(q, k) \right\}. \end{aligned} \quad (39)$$

Then, similarly as in Theorem 1,

$$\mathcal{R}_q(x^n) \subseteq \mathcal{R}_\circ^k(x^n) \triangleq \bigcup_{(\hat{x}^n, \tilde{x}^n) \in \mathcal{B}(x^n)} \mathcal{R}_-^k(\hat{x}^n, \tilde{x}^n), \quad (40)$$

and so, for every positive integer  $N$ :

$$\bigcap_{n \geq N} \mathcal{R}_q(x^n) \subseteq \bigcap_{n \geq N} \mathcal{R}_\circ^k(x^n), \quad (41)$$

implying that

$$\mathcal{R}_q(\mathbf{x}) = \bigcup_{N \geq 1} \bigcap_{n \geq N} \mathcal{R}_q(x^n) \subseteq \bigcup_{N \geq 1} \bigcap_{n \geq N} \mathcal{R}_\circ^k(x^n) \triangleq \mathcal{R}_\circ^k(\mathbf{x}). \quad (42)$$

Since this holds for every positive integer  $k$ , then

$$\mathcal{R}_q(\mathbf{x}) \subseteq \bigcup_{K \geq 1} \bigcap_{k \geq K} \mathcal{R}_o^k(\mathbf{x}) \triangleq \mathcal{R}_o(\mathbf{x}), \quad (43)$$

and so,

$$\mathcal{R}_\infty(\mathbf{x}) \subseteq \mathcal{R}_o(\mathbf{x}), \quad (44)$$

which establishes an asymptotic version of the converse theorem.

As for the direct part, considering the fact that a block code of length  $k$ , operating on  $k$ -tuples of the two reconstruction vectors can be implemented by a finite-state machine with no more than  $(\beta\gamma)^k$  states, we have

$$\begin{aligned} \mathcal{R}_{(\beta\gamma)^k}(x^n) &\supseteq \mathcal{R}_i^k(x^n) \\ &\triangleq \bigcup_{(\hat{x}^n, \tilde{x}^n) \in \mathcal{B}(x^n)} \mathcal{R}_+^k(\hat{x}^n, \tilde{x}^n), \end{aligned} \quad (45)$$

where

$$\begin{aligned} \mathcal{R}_+^k(\hat{x}^n, \tilde{x}^n) &= \left\{ (R_1, R_2) : R_1 \geq \frac{k}{n} \sum_{t=0}^{n/k-1} \rho_{\text{LZ}}(\hat{x}_{kt+1}^{kt+k}) + O\left(\frac{1}{\log k}\right), \right. \\ &\quad R_1 + R_2 \geq \frac{k}{n} \sum_{t=0}^{n/k-1} \left[ \rho_{\text{LZ}}(\hat{x}_{kt+1}^{kt+k}) + \right. \\ &\quad \left. \left. \rho_{\text{LZ}}(\tilde{x}_{kt+1}^{kt+k} | \hat{x}_{kt+1}^{kt+k}) \right] + O\left(\frac{\log(\log k)}{\log k}\right) \right\}. \end{aligned} \quad (46)$$

$$\begin{aligned} \mathcal{R}_{(\beta\gamma)^k}(\mathbf{x}) &= \bigcup_{N \geq 1} \bigcap_{n \geq N} \mathcal{R}_{(\beta\gamma)^k}(x^n) \\ &\supseteq \bigcup_{N \geq 1} \bigcap_{n \geq N} \mathcal{R}_i^k(x^n) \\ &= \mathcal{R}_i^k(\mathbf{x}) \\ &\supseteq \bigcap_{K \geq k} \mathcal{R}_i^K(\mathbf{x}) \end{aligned} \quad (47)$$

and so,

$$\mathcal{R}_\infty(\mathbf{x}) = \bigcup_{k \geq 1} \mathcal{R}_{(\beta\gamma)^k}(\mathbf{x}) \supseteq \bigcup_{k \geq 1} \bigcap_{K \geq k} \mathcal{R}_i^K(\mathbf{x}) = \mathcal{R}_i(\mathbf{x}). \quad (48)$$

We have just proved the following theorem:

**Theorem 2.** For every infinite individual sequence  $\mathbf{x} = (x_1, x_2, \dots)$ ,

$$\mathcal{R}_o(\mathbf{x}) \supseteq \mathcal{R}_\infty(\mathbf{x}) \supseteq \mathcal{R}_i(\mathbf{x}). \quad (49)$$

These inner and outer bounds are tight in the sense that the definitions of  $\mathcal{R}_o(\mathbf{x})$  and  $\mathcal{R}_i(\mathbf{x})$  are based on the same building blocks and the only difference is in terms that tend to zero as  $k \rightarrow \infty$ .

## 5 Multiple Description Coding

Consider next the configuration that is associated with the multiple description problem (see, e.g., Chap. 13 of [9]), where the source is an individual sequence and the encoders are modeled as finite-state machines. In particular, there are two  $q$ -state encoders and three decoders defined as follows. Encoders 1 and 2 are fed by  $x^n$  and produce two reconstructions,  $\hat{x}^n$  and  $\tilde{x}^n$ , with distortions  $d_1(x^n, \hat{x}^n) \leq nD_1$  and  $d_2(x^n, \tilde{x}^n) \leq nD_2$ , respectively. Encoder 1 then compresses  $\hat{x}^n$  losslessly and sends a compressed description to Decoder 1. Likewise, Encoder 2 does the same with  $\tilde{x}^n$  and sends a compressed form to Decoder 2. There is no collaboration between Decoders 1 and 2. The third decoder, Decoder 0, receives both compressed descriptions and generates yet another reconstruction,  $\dot{x}^n$ , with distortion  $d_0(x^n, \dot{x}^n) \leq nD_0$ .

Using the same technique as in the proof of Theorem 1, it is easy to prove the following outer bound to the achievable rate region:

$$\mathcal{R}_o(x^n) = \bigcup_{(\hat{x}^n, \tilde{x}^n, \dot{x}^n) \in \mathcal{B}(x^n)} \mathcal{R}_{\text{LZ}}(\hat{x}^n, \tilde{x}^n, \dot{x}^n), \quad (50)$$

where  $\mathcal{B}(x^n)$  is redefined as

$$\mathcal{B}(x^n) = \{(\hat{x}^n, \tilde{x}^n, \dot{x}^n) : d_0(x^n \dot{x}^n) \leq nD_0, d_1(x^n \hat{x}^n) \leq nD_1, d_2(x^n \tilde{x}^n) \leq nD_2\}, \quad (51)$$

and

$$\begin{aligned} \mathcal{R}_{\text{LZ}}(\hat{x}^n, \tilde{x}^n, \dot{x}^n) = & \left\{ (R_1, R_2) : R_1 \geq \rho_{\text{LZ}}(\hat{x}^n) - \Delta_1(q, n), R_2 \geq \rho_{\text{LZ}}(\tilde{x}^n) - \Delta_1(q, n), \right. \\ & R_1 + R_2 \geq \rho_{\text{LZ}}(\dot{x}^n | \hat{x}^n, \tilde{x}^n) + \\ & \left. \rho_{\text{LZ}}(\hat{x}^n, \tilde{x}^n) - \Delta_2(q, n) \right\}, \end{aligned} \quad (52)$$

with  $\Delta_1(q, n)$  and  $\Delta_2(q, n)$  being defined similarly as before. The sum-rate inequality is obtained by considering that the two encoders together compress losslessly the triple  $(\hat{x}^n, \tilde{x}^n, \dot{x}^n)$  and so, the main term of the lower bound to  $R_1 + R_2$  is the joint empirical entropy of  $(\hat{x}^n, \tilde{x}^n, \dot{x}^n)$ , which can be decomposed as the sum of the joint empirical entropy of  $(\hat{x}^n, \tilde{x}^n)$  and the conditional empirical entropy of  $\dot{x}^n$  given  $(\hat{x}^n, \tilde{x}^n)$ , which in turn are essentially further lower bounded by  $\rho_{\text{LZ}}(\hat{x}^n, \tilde{x}^n)$  and  $\rho_{\text{LZ}}(\dot{x}^n | \hat{x}^n, \tilde{x}^n)$ , respectively.

We next present two inner bounds, where the first one is analogous to the El Gamal-Cover inner bound [20] and the second follows the same line of thought as that of the Zhang-Berger inner bound [21].

The former inner bound is given by

$$\mathcal{R}_i^{\text{EGC}}(x^n) = \bigcup_{(\hat{x}^n, \tilde{x}^n, \dot{x}^n) \in \mathcal{B}(x^n)} \mathcal{R}_i(\hat{x}^n, \tilde{x}^n, \dot{x}^n), \quad (53)$$

where

$$\begin{aligned} \mathcal{R}_i(\hat{x}^n, \tilde{x}^n, \dot{x}^n) = & \left\{ (R_1, R_2) : R_1 \geq \rho_{\text{LZ}}(\hat{x}^n) + \epsilon(n), R_2 \geq \rho_{\text{LZ}}(\tilde{x}^n) + \epsilon(n), \right. \\ & \left. R_1 + R_2 \geq \rho_{\text{LZ}}(\dot{x}^n | \hat{x}^n, \tilde{x}^n) + \rho_{\text{LZ}}(\hat{x}^n, \tilde{x}^n) + \hat{I}(\hat{x}^n; \tilde{x}^n) + \epsilon(n) + \hat{\epsilon}(n) \right\}, \end{aligned} \quad (54)$$

where  $\epsilon(n)$  and  $\hat{\epsilon}(n)$  are as in (13) and (18), respectively.

$$\hat{I}(\hat{x}^n; \tilde{x}^n) = \rho_{\text{LZ}}(\hat{x}^n) + \rho_{\text{LZ}}(\tilde{x}^n) - \rho_{\text{LZ}}(\hat{x}^n, \tilde{x}^n). \quad (55)$$

The quantity  $\hat{I}(\hat{x}^n; \tilde{x}^n)$  plays a role of an empirical mutual information between  $\hat{x}^n$  and  $\tilde{x}^n$ , which manifests the gap between the lower bounds to the sum-rate inequalities of the inner bound and the outer bound, analogously to the mutual information term of the El Gamal-Cover achievable region.

The achievability of the above inner bound is as follows. Given an internal point in  $\mathcal{R}_i^{\text{EGC}}(x^n)$ , there must exist a reconstruction triple  $(\hat{x}^n, \tilde{x}^n, \hat{x}^n)$  that meets the distortion constraints and the corresponding rate inequalities. The encoder applies individual LZ compression for both  $\hat{x}^n$  and  $\tilde{x}^n$  and sends the compressed versions, at rates  $\rho_{\text{LZ}}(\hat{x}^n)$  and  $\rho_{\text{LZ}}(\tilde{x}^n)$  (up to negligibly small terms for large  $n$ ), to Decoder 1 and Decoder 2, respectively. It then applies conditional LZ compression of  $\hat{x}^n$  given  $(\hat{x}^n, \tilde{x}^n)$  at rate  $\rho_{\text{LZ}}(\hat{x}^n | \hat{x}^n, \tilde{x}^n)$  (up to small terms), and splits this compressed bit stream between Decoders 1 and 2 without violating their rate inequalities. The rate sum is then essentially

$$\begin{aligned} & \rho_{\text{LZ}}(\hat{x}^n | \hat{x}^n, \tilde{x}^n) + \rho_{\text{LZ}}(\hat{x}^n) + \rho_{\text{LZ}}(\tilde{x}^n) \\ = & \rho_{\text{LZ}}(\hat{x}^n | \hat{x}^n, \tilde{x}^n) + \rho_{\text{LZ}}(\hat{x}^n, \tilde{x}^n) + \hat{I}(\hat{x}^n; \tilde{x}^n). \end{aligned} \quad (56)$$

In the above description, we explained that the bit-stream associated with the conditional compression of  $\hat{x}^n$  given  $(\hat{x}^n, \tilde{x}^n)$  is split between Decoders 1 and 2 without violating the rate inequalities. This is always possible because of the following simple fact: Given an internal point in the region  $\mathcal{R} = \{(R_1, R_2) : R_1 > A, R_2 > B, R_1 + R_2 \geq A + B + C\}$ , there must exist  $0 \leq D \leq C$  such that  $(A + D, B + C - D) \in \mathcal{R}$ . In particular, let  $D = R_1 - A \geq 0$ . Then,  $R_2 = B + C - D \geq B$  and  $R_1 + R_2 = (A + D) + (B + C - D) = A + B + C$ . In our case,  $A = \rho_{\text{LZ}}(\hat{x}^n)$ ,  $B = \rho_{\text{LZ}}(\tilde{x}^n)$ , and  $C = \rho_{\text{LZ}}(\hat{x}^n | \hat{x}^n, \tilde{x}^n)$ .

The second achievability scheme, in the spirit of the Zhang-Berger scheme, is as follows. Here, in addition to  $\hat{x}^n$ ,  $\tilde{x}^n$  and  $\hat{x}^n$ , we also generate an auxiliary finite-alphabet sequence,  $u^n$ . The encoder applies LZ compression to  $u^n$  and conditional LZ compression of  $\hat{x}^n$  given  $u^n$  and sends both bit-streams to Decoder 1. At the same time, it also applies conditional LZ compression of  $\tilde{x}^n$  given  $u^n$  and sends the compressed forms of  $u^n$  and  $\tilde{x}^n$  to Decoder 2. Finally, the encoder applies conditional LZ compression of  $\hat{x}^n$  given  $(\hat{x}^n, \tilde{x}^n, u^n)$  and splits the compressed bit-stream between Decoders and Decoder 2 in a manner that meets the rate constraints. Thus,

$$R_1 \approx \rho_{\text{LZ}}(u^n) + \rho_{\text{LZ}}(\hat{x}^n|u^n) + \alpha\rho_{\text{LZ}}(\dot{x}^n|\hat{x}^n, \tilde{x}^n, u^n) \quad (57)$$

$$R_2 \approx \rho_{\text{LZ}}(u^n) + \rho_{\text{LZ}}(\tilde{x}^n|u^n) + (1 - \alpha)\rho_{\text{LZ}}(\dot{x}^n|\hat{x}^n, \tilde{x}^n, u^n), \quad (58)$$

where  $\alpha \in [0, 1]$ . Thus,

$$\begin{aligned} R_1 + R_2 &\approx 2 \cdot \rho_{\text{LZ}}(u^n) + \rho_{\text{LZ}}(\hat{x}^n|u^n) + \rho_{\text{LZ}}(\tilde{x}^n|u^n) + \rho_{\text{LZ}}(\dot{x}^n|\hat{x}^n, \tilde{x}^n, u^n) \\ &= 2 \cdot \rho_{\text{LZ}}(u^n) + \rho_{\text{LZ}}(\hat{x}^n, \tilde{x}^n|u^n) + \rho_{\text{LZ}}(\dot{x}^n|\hat{x}^n, \tilde{x}^n, u^n) + \hat{I}(\hat{x}^n; \tilde{x}^n|u^n), \end{aligned} \quad (59)$$

where

$$\hat{I}(\hat{x}^n; \tilde{x}^n|u^n) = \rho_{\text{LZ}}(\hat{x}^n|u^n) + \rho_{\text{LZ}}(\tilde{x}^n|u^n) - \rho_{\text{LZ}}(\hat{x}^n, \tilde{x}^n|u^n), \quad (60)$$

is analogous to conditional mutual information. The first term in the rate sum is analogous to  $2I(U; X)$ , the sum of the second and the third is analogous to  $I(X; \hat{X}, \tilde{X}, \dot{X}|U)$  and the last term is analogous to  $I(\hat{X}; \tilde{X}|U)$  (see Theorem 13.4, page 332 in [9]).

## References

- [1] Equitz, W. H. R.; Cover T. M. “Successive refinement of information,” *IEEE Trans. Inform. Theory*, **1991**, vol. IT-37, no. 2, pp. 269–275.
- [2] Koshelev, V. N. “On the divisibility of discrete sources with an additive single-letter distortion measure,” *Problems of Information Transmission (IPPI)*, **1994**, vol. 30, no. 1, pp. 27–43.
- [3] Rimoldi, B, “Successive refinement of information: characterization of achievable rates,” *IEEE Trans. Inform. Theory*, **1994**, vol. IT-40, no. 1, pp. 253–259.
- [4] Kanlis, A.; Narayan, P. “Error exponents for successive refinement by partitioning,” *IEEE Trans. Inform. Theory*, **1996**, vol. IT-42, no. 1.
- [5] Steinberg, Y.; Merhav, N. “On successive refinement for the Wyner–Ziv problem,” *IEEE Trans. Inform. Theory*, **2004**, vol. 50, no. 8, pp. 1636–1654.

- [6] Maor, A.; Merhav, N. “On successive refinement with causal side information at the decoders,” *IEEE Trans. Inform. Theory*, **2008**, vol. 54, no. 1, pp. 332–343.
- [7] Maor (Varshavsky), A; Merhav, N. “On successive refinement for the Kaspi/Heegard–Berger problem,” *IEEE Trans. Inform. Theory*, **2010**, vol. 56, no. 8, pp. 3930–3945.
- [8] Tian, C.; Diggavi, S. N. “Multistage successive refinement for Wyner-Ziv source coding with degraded side informations,” *Proc. 2006 IEEE International Symposium on Information Theory (ISIT 2006)*, Seattle, WA, USA, 2006, pp. 1594–1598, doi: 10.1109/ISIT.2006.261545.
- [9] El Gamal, A.; Kim, Y.-H. *Network Information Theory*, Cambridge University Press, Cambridge, UK, 2011.
- [10] Lempel, A.; Ziv, J. “On the complexity of finite sequences,” *IEEE Trans. Inform. Theory*, **1976** vol. IT-22, no. 1, pp. 137–143.
- [11] Ziv, J.; Lempel, A. “Compression of individual sequences via variable-rate coding,” *IEEE Trans. Inform. Theory* **1978**, vol. IT-24, no. 5, pp. 530–536.
- [12] Ziv, J. “Distortion-rate theory for individual sequences,” *IEEE Trans. Inform. Theory*, **1980** vol. IT-26, no. 2, pp. 137–143.
- [13] Ziv, J. “Fixed-rate encoding of individual sequences with side information,” *IEEE Trans. Inform. Theory*, **1984**, vol. IT-30, no. 2, pp. 348–352.
- [14] Yang, E.-h.; Kieffer, J. C. “Simple universal lossy data compression schemes derived from the Lempel-Ziv algorithm,” *IEEE Trans. Inform. Theory* **1996**, vol. 42, no. 1, pp. 239–245.
- [15] Uyematsu, T.; Kuzuoka, S, “Conditional Lempel-Ziv complexity and its application to source coding theorem with side information,” *IEICE Trans. Fundamentals* **2003**, Vol. E86-A, no. 10, pp. 2615–2617.
- [16] Merhav, N. “Universal detection of messages via finite-state channels,” *IEEE Trans. Inform. Theory* **2000**, vol. 46, no. 6, pp. 2242–2246.
- [17] Merhav, N. “A universal random coding ensemble for sample-wise lossy compression,” *Entropy*, **2023**, 25(8), 1199; <https://doi.org/10.3390/e25081199>, August 2023.

- [18] Merhav, N. “Lossy compression of individual sequences revisited: fundamental limits of finite-state encoders,” *Entropy* **2024**, 26(2), 116.  
<https://doi.org/10.3390/e26020116> January 28, 2024.
- [19] Merhav, N. “Universal Slepian-Wolf coding for individual sequences,” *IEEE Trans. Inform. Theory* **2025**, vol. 71, no. 1, pp. 783–796.
- [20] El Gamal, A.; Cover, T. M. “Achievable rates for multiple descriptions,” *IEEE Trans. Inform. Theory*, **1982**, vol. 28, no. 6, pp. 851–857.
- [21] Zhang, Z.; Berger, T. “New results in binary multiple descriptions,” *IEEE Trans. Inform. Theory*, **1987**, vol. 33, no. 4, pp. 502–521.
- [22] Merhav, N. “A universal random coding ensemble for sample-wise lossy compression,” *Entropy* 2023, 25, 1199.
- [23] Ziv, J. “Universal decoding for finite-state channels,” *IEEE Trans. Inform. Theory*, **1985**, vol. IT–31, no. 4, pp. 453–460.