On the Wyner–Ziv Problem for Individual Sequences

Neri Merhav and Jacob Ziv

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, ISRAEL

February 10, 2005

Abstract

We consider a variation of the Wyner–Ziv problem pertaining to lossy compression of individual sequences using finite–state encoders and decoders. There are two main results in this paper. The first characterizes the relationship between the performance of the best M–state encoder–decoder pair to that of the best block code of size ℓ for every input sequence, and shows that the loss of the latter relative to the former (in terms of both rate and distortion) never exceeds the order of $(\log M)/\ell$, independently of the input sequence. Thus, in the limit of large M, the best rate–distortion performance of every infinite source sequence can be approached universally by a sequence of block codes (which are also implementable by finite–state machines). While this result assumes an asymptotic regime where the number of states is fixed, and only the length n of the input sequence grows without bound, we then consider the case where the number of states $M = M_n$ is allowed to grow concurrently with n. Our second result is then about the critical growth rate of M_n such that the rate–distortion performance of M_n–state encoder–decoder pairs can still be matched by a universal code. We show that this critical growth rate is of M_n is linear in n.

Index Terms: Finite–state machines, individual sequences, side information, block codes, universal coding, Wyner–Ziv problem.

1 Introduction

In a series of papers from the late seventies until the mid–eighties, Ziv [11],[12],[13], and Ziv and Lempel [14],[4] have developed a theory of universal compression of individual sequences using finite–state machines (FSM's). In particular, the work [11] focuses on universal, fixed–rate, (almost) lossless compression of individual sequences using finite–state encoders and decoders, which was then further developed to the well–known Lempel–Ziv algorithm [14],[4]. In [12], the framework of [11] was extended to lossy compression, and in [13], the results of [11] were extended in another direction, pertaining to almost lossless compression

in the presence of an (individual) side information sequence at the decoder, namely, an analogue to Slepian–Wolf coding [7] for individual sequences.

In this work, we take yet another step in this direction and further generalize this model setting, of universal coding for individual sequences using finite–state encoders and decoders, to that of lossy compression in the presence of side information at the decoder, in other words, Wyner–Ziv (W–Z) coding [10] for individual sequences. Also, unlike the fixed–rate codes assumed in [11],[12],[13], here our model allows variable–rate coding, which give rise to considerably more flexibility. On the other hand, in our model, the side information sequence at the decoder is assumed to be generated from the individual source sequence (to be compressed) via a known memoryless channel, in contrast to [13], where it is modelled as another individual sequence.[1] Furthermore, our model setting can also be viewed as an extension of the setting of universal finite–state denoising of individual sequences corrupted by stochastic noise (cf. [9], [6] and references therein): The denoising problem is actually a special case of this W–Z model, where the coding rate is zero.

There are two main results in this paper. The first result is a characterization of the relationship between the performance of the best M–state finite–state encoder–decoder pair (for a given input sequence) to that of the best block code of size ℓ for every input sequence, and it shows that the loss of the latter relative to the former (in terms of both rate and distortion) is of the order of $(\log M)/\ell$, independently of the input sequence. Thus, in the limit of large M, the best rate–distortion performance of every infinite source sequence can be approached universally by a sequence of block codes (which are also implementable by finite–state machines). One of the interesting features of these universal codes is that they require no binning, as opposed to the well–known W–Z code in classical in the probabilistic case [10]. We also extend this result to framework of successive refinement coding (cf. e.g., [2], [5], [8]), where there are two encoders and two decoders (all of which are finite–state machines): The first encoder transmits a relatively coarse description to the first decoder, which has access also to a certain side information stream. The second encoder sends a refinement code, to another decoder that has access also to the first compressed bitstream, as well as to another side information sequence. This is setup is in analogy to a recent study

[1] The reason for this assumption is that even in the classical, probabilistic setting, W–Z coding cannot be universal w.r.t. the channel from the source to the side information stream (unless there is feedback) as the encoder, which has no access to the side information, cannot 'learn' its statistics. This is different from the Slepian–Wolf setting, where the encoder does not depend on these statistics.

on successive refinement coding for the W–Z problem for probabilistic memoryless sources [8]. However, in contrast to [8], where a certain Markov structure had to be assumed regarding the source and the two side information processes, here no such structure is needed. Also, unlike in [8], here the extension to multiple stages is straghtforward.

Returning to the single–stage coding model, we next relax the assumption that the number of states is fixed, independently of the length n of the input data. In other words, we examine an asymptotic regime that allows the number of states $M = M_n$ to grow concurrently with n, and we investigate the critical growth rate of M_n such that the rate–distortion performance of M_n–state encoder–decoder pairs can still be matched by a universal code (in a sense to be made precise later on). Our second result is that this critical growth rate of M_n is linear in n, in the sense that if $M_n = n^\theta$, universal achievability is guaranteed for all $\theta < 1$, but not for $\theta > 1$. In other words, $\theta = 1$ is the critical value of θ.

In this context, it is interesting to go back, for a moment, to the lossless case without side information and to consider the performance of the well–known Lempel–Ziv algorithm in that respect. By examining the converse to the coding theorem (Theorem 1) in [14], which states that the best compression achievable by an FSM with M states is lower bounded by a quantity the behaves roughly like $c\log(c/4M^2)$ where c is the number of distinct phrases in the input sequence. Since the length of the LZ code is about $c\log c$, the gap, $c\log(4M^2)$, would be relatively negligible only as long as $\log M$ would be very small compared to $\log c \sim \emptyset(\log n)$, which is guarateed to be the case when θ is very small ($\theta << 1$). It, therefore, turns out that there is a gap between the best that can be done by a universal code, where θ can be chosen arbitrarily clode to unity, and the performance of the LZ algorithm in that respect.

The outline of this paper is as follows. In Section 2, we define the problem and establish notation conventions. Section 3 is devoted to the derivation of the first result described above, pertaining to a fixed number of states. Finally, Section 4 focuses on the critical growth rate of the number of states. The extension of the results of Section 3 to successive refinement coding is deferred to the Appendix, for the sake of continuity between Sections 3 and 4 which are both about single–stage codes.

2 Notation and Problem Formuation

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, as well as some other sets, will be denoted by calligraphic letters. Similarly, random vectors, their realizations, and their alphabets, will be denoted, respectively, by capital letters, the corresponding lower case letters, and calligraphic letters, all superscripted by their dimensions. For example, the random vector $Y^n = (Y_1, \ldots, Y_n)$, (n – positive integer) may take a specific vector value $y^n = (y_1, \ldots, y_n)$ in $\mathcal{Y}^n$, the nth order Cartesian power of $\mathcal{Y}$, which is the alphabet of each component of this vector. For $i \leq j$ (i, j – positive integers), x_i^j will denote the segment $(x_i, \ldots, x_j)$, where for $i = 1$ the subscript will be omitted.

Let $\boldsymbol{x} = (x_1, x_2, \ldots)$, $x_i \in \mathcal{X}$, $i = 1, 2, \ldots$, with $|\mathcal{X}| = \alpha < \infty$, be an infinite input sequence (to be compressed) and let $\boldsymbol{y} = (y_1, y_2, \ldots)$, $y_i \in \mathcal{Y}$, $i = 1, 2, \ldots$, with $|\mathcal{Y}| = \beta < \infty$, be a corresponding side–information sequence generated by a given discrete memoryless channel (DMC)

$$P(y_1, \ldots, y_n | x_1, \ldots, x_n) = \prod_{i=1}^{n} P(y_i | x_i), \quad n = 1, 2, \ldots. \tag{1}$$

When the sequence $\boldsymbol{x}$ is sequentially fed into a variable–rate finite–state encoder $\mathcal{E} = (\mathcal{S}, f, g)$, this encoder generates an infinite sequence of binary strings of variable length, $\boldsymbol{u} = (u_1, u_2, \ldots)$, while going through an infinite sequence of states $s_1, s_2, \ldots$ according to

$$\begin{aligned} u_i &= f(s_i, x_i), \\ s_{i+1} &= g(s_i, x_i), \quad i = 1, 2, \ldots \end{aligned} \tag{2}$$

where the initial state s_1 is assumed to be a certain fixed member of $\mathcal{S}$. At the same time and in a similar manner, the finite–state decoder $\mathcal{D} = (\mathcal{S}', f', g')$ sequentially maps $\boldsymbol{u}$ and $\boldsymbol{y}$ to an infinite reproduction sequence $\hat{x}_1, \hat{x}_2, \ldots$, $\hat{x}_i \in \hat{\mathcal{X}}$, $i = 1, 2, \ldots$, with $|\hat{\mathcal{X}}| = \gamma < \infty$, using the recursion

$$\begin{aligned} \hat{x}_{i-d} &= f'(s'_i, u_i, y_i), \quad i = d+1, d+2, \ldots \\ s'_{i+1} &= g'(s'_i, u_i, y_i), \quad i = 1, 2, \ldots. \end{aligned} \tag{3}$$

where d (positive integer) is the encoding–decoding delay, and the initial state s'_1 is assumed to be a certain fixed member of $\mathcal{S}'$. It is assumed that at each time instant i, when the

decoder is at state s'_i, it is able to isolate the current input codeword u_i from the following codewords of the compressed bitstream, $u_{i+1}, u_{i+2}, \ldots$.[2] To this end, we allow a prefix code $\mathcal{C}(s')$ associated with each $s' \in \mathcal{S}'$ (with the option that for some $s' \in \mathcal{S}'$, $\mathcal{C}(s')$ may be empty, in which case, the decoder idles). For every $u \in \mathcal{C}(s')$, let $L(u)$ denote the length of u (in bits). We then assume that the Kraft inequality

$$\sum_{u \in \mathcal{C}(s')} 2^{-L(u)} \leq 1 \tag{4}$$

holds for all $s' \in \mathcal{S}'$. Note that the above discussion continues to apply when single codewords are replaced by ℓ–vectors, u^ℓ, formed by concatenating ℓ (legitimate) codewords successively. In this case, let $\mathcal{C}^\ell(s')$ denote the supercode formed by all $\{u^\ell\}$ that originate from state s'. Clearly, since the components of u^ℓ can be identified recursively, the supercode satisfies the Kraft inequality as well w.r.t. the length function $L(u^\ell) = \sum_{i=1}^{\ell} L(u_i)$.

For a given single–letter distortion measure $\rho : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$, let $\Delta(x^n, \mathcal{E}, \mathcal{D})$ denote the expected distortion $\frac{1}{n}\sum_{i=1}^{n} E\rho(x_i, \hat{X}_i)$ associated with the encoding and decoding x^n by $(\mathcal{E}, \mathcal{D})$, where the expectation is w.r.t. the DMC. Now, define

$$\Delta_{M,d}(x^n, R) = \min \Delta(x^n, \mathcal{E}, \mathcal{D}) \tag{5}$$

where the minimum is over all encoder–decoder pairs $\{(\mathcal{E}, \mathcal{D})\}$ having no more than M states each, with delay no longer than d, and which satisfy the rate constraint:

$$\sum_{i=1}^{n} L(u_i) \leq nR. \tag{6}$$

While the optimum $(\mathcal{E}, \mathcal{D})$ for achieving $\Delta_{M,d}(x^n, R)$ depends, in general, on x^n, we are interested in a universal algorithm (independent of x^n) that 'competes' with the best M–state encoder–decoder pair $(\mathcal{E}, \mathcal{D})$, and eventually approaches the *operational* (finite–state) W–Z distortion–rate function, which we define as:

$$\Delta(\boldsymbol{x}, R) \stackrel{\Delta}{=} \lim_{d \to \infty} \lim_{M \to \infty} \limsup_{n \to \infty} \Delta_{M,d}(x^n, R). \tag{7}$$

Note that the order of the limits is first over M and then over d. This is because given M, the delay d cannot be arbitrarily large. To implement a finite–state machine with delay d, such a system must store the d most recent inputs, which requires the number of states to be

[2] To this end, it would make sense to assume that the dependence of u_i on s'_i is only via a part of the decoder state information that is independent of the (random) SI sequence, e.g., s''_i, which is updated according to $s''_{i+1} = g''(s''_i, u_i)$.

at least exponential in d, namely, the maximum possible delay for a given M, which we shall denote by d_M, is proportional to $\log M$. A somewhat stronger notion of the operational finite–state W–Z distortion–rate function is then given by

$$\Delta^*(\boldsymbol{x}, R) \stackrel{\Delta}{=} \lim_{M\to\infty} \limsup_{n\to\infty} \Delta_{M,d_M}(x^n, R). \tag{8}$$

Obviously, $\Delta^*(\boldsymbol{x}, R) \le \Delta(\boldsymbol{x}, R)$.

3 Long Block Codes are as Good as FSM's

We start by defining the *informational* W–Z distortion–rate function (as opposed to the operational definitions of eqs. (7), (8)) in the following manner: Let n and $\ell < n$ be two given positive integers, assume, without essential loss of generality, that ℓ divides n, and let us chop the sequence x^n into ℓ–blocks, $\{x_{i\ell+1}^{(i+1)\ell}\}_{i=0}^{n/\ell-1}$. Now, let $P(a^\ell)$, $a^\ell = (a_1, \ldots, a_\ell) \in \mathcal{X}^\ell$, denote the empirical probability (relative frequency) of the ℓ–vector a^ℓ along x^n, i.e.,

$$P(a^l) = \frac{\ell}{n} \sum_{k=0}^{n/\ell-1} 1\{x_{i\ell+1}^{(i+1)\ell} = a^\ell\} \tag{9}$$

and for any $b^\ell = (b_1, \ldots, b_\ell) \in \mathcal{Y}^\ell$, let

$$P(a^\ell, b^\ell) = P(a^\ell)P(b^\ell|a^\ell) = P(a^\ell) \cdot \prod_{i=1}^{\ell} P(b_i|a_i), \tag{10}$$

where $\{P(b_i|a_i)\}$ are the single–letter transition probabilities associated with the DMC (1). Let (X^ℓ, Y^ℓ) designate random ℓ–vectors jointly distributed according to $\{P(a^\ell, b^\ell),\ a^\ell \in \mathcal{X}^\ell,\ b^\ell \in \mathcal{Y}^\ell\}$, and define the ℓ–th order informational W–Z distortion–rate function of the source X^ℓ w.r.t. the SI Y^ℓ as follows:

$$\Delta_{X^\ell|Y^\ell}(R) = \min_{U,h} \frac{1}{\ell} E\rho(X^\ell, h(Y^\ell, U)) \tag{11}$$

where the minimum is over all functions $\{h\}$ from $\mathcal{Y}^\ell \times \mathcal{U}$ to $\hat{\mathcal{X}}^\ell$, and over all RV's $\{U\}$ that: (i) take on values in an alphabet $\mathcal{U}$ of size $|\mathcal{X}|^\ell + 1$, (ii) satisfy the Markov relation $U \to X^\ell \to Y^\ell$, and (iii) satisfy the inequality

$$H(U) \le \ell R. \tag{12}$$

Our main result in this section relates the operational W–Z distortion–rate function, for finite M and d, to the ℓ–th order informational distortion–rate function associated with the empirical distribution of x^n, as defined above:

Theorem 1 *For every positive integers n and ℓ such that ℓ divides n:*

$$\Delta_{M,d}(x^n, R) \geq \Delta_{X^\ell|Y^\ell}\left(R + \frac{2\log M}{\ell}\right) - \frac{\rho_{\max} d}{\ell}, \tag{13}$$

where $\rho_{\max} \stackrel{\Delta}{=} \max_{x,\hat{x}} \rho(x, \hat{x})$.

Note that the rate–redundancy, $(2\log M)/\ell$, depends on the number of states whereas the distortion redundancy, $\rho_{\max} d/\ell$, depends only the delay. However, referring to relationship between d and M (cf. the last paragraph of Section 2), the distortion–redundancy term $\rho_{\max} d/\ell$, is also bounded by a quantity proportional to $(\log M)/\ell$, like the rate–redundancy.

Defining now the informational W–Z distortion–rate function of the infinite sequence $\boldsymbol{x}$ as

$$\Delta_{\boldsymbol{X}|\boldsymbol{Y}}(R) \stackrel{\Delta}{=} \limsup_{\ell\to\infty} \limsup_{n\to\infty} \Delta_{X^\ell|Y^\ell}(R), \tag{14}$$

we have the following corollary to Theorem 1, which means that long block codes are asymptotically sufficiently good to attain the asymptotic performance of general finite–state codes:

Corollary 1 *For every $\boldsymbol{x}$,*

$$\Delta_{\boldsymbol{X}|\boldsymbol{Y}}(R) \geq \Delta^*(\boldsymbol{x}, R) \geq \Delta_{\boldsymbol{X}|\boldsymbol{Y}}(R^+) \stackrel{\Delta}{=} \lim_{R'\downarrow R} \Delta_{\boldsymbol{X}|\boldsymbol{Y}}(R'). \tag{15}$$

Observe that, by monotonicity, $\Delta_{\boldsymbol{X}|\boldsymbol{Y}}(R^+) = \Delta_{\boldsymbol{X}|\boldsymbol{Y}}(R)$ (hence the two inequalities become equalities) for every $R \geq 0$, with the possible exception of a countable set of points. Corollary 1 follows from Theorem 1 in a simple manner: For any given $R' > R$, Theorem 1 implies that $\Delta_{M,d_M}(x^n, R) \geq \Delta_{X^\ell|Y^\ell}(R') - \rho_{\max} d_M/\ell$ for all sufficiently large ℓ. Taking first, the limsup as $n \to \infty$, next the limsup as $\ell \to \infty$, then the limit as $M \to \infty$, and finally, the limit $R' \to R$, gives the right inequality. The left inequality follows from the fact that a block code of length ℓ is implementable by an FSM with α^ℓ states (cf. [11, Example 2, p. 406]) thus $\Delta_{X^\ell|Y^\ell}(R) \geq \Delta_{\alpha^\ell, d_{\alpha^\ell}}(x^n, R)$, and the result is obtained by taking first the limsup over n and then the limsup over ℓ at both sides of this inequality.

We now turn to the proof of Theorem 1.

Proof. For a given combination of an M–state encoder and an M–state decoder $(\mathcal{E}, \mathcal{D})$, consider the joint probability distribution

$$P(a^\ell, c^\ell, s, s') = \frac{\ell}{n} \sum_{i=0}^{n/\ell-1} 1\{x_{i\ell+1}^{i\ell+\ell} = a^\ell, u_{i\ell+1}^{i\ell+\ell} = c^\ell, s_{i\ell+1} = s, s'_{i\ell+1} = s'\}, \tag{16}$$

for every $a^\ell \in \mathcal{X}^\ell$, $s \in \mathcal{S}$, $s' \in \mathcal{S}'$, and $c^\ell \in \mathcal{C}^\ell(s')$. Note that $\hat{x}_{i\ell-d+1}^{i\ell-d+\ell}$ depends (deterministically) only on $y_{i\ell+1}^{i\ell+\ell}$, $u_{i\ell+1}^{i\ell+\ell}$, and $s'_{i\ell+1}$. Let us denote then $\hat{x}_{i\ell-d+1}^{i\ell-d+\ell} = h(y_{i\ell+1}^{i\ell+\ell}, u_{i\ell+1}^{i\ell+\ell}, s'_{i\ell+1})$, and assuming that $\ell > d$, let $\hat{x}_{i\ell+1}^{i\ell-d+\ell} \stackrel{\Delta}{=} h'(y_{i\ell+1}^{i\ell+\ell}, u_{i\ell+1}^{i\ell+\ell}, s'_{i\ell+1})$ be defined simply by truncating the first d components of $h(y_{i\ell+1}^{i\ell+\ell}, u_{i\ell+1}^{i\ell+\ell}, s'_{i\ell+1})$. Now, define

$$\begin{aligned} P(a^\ell, b^\ell, c^\ell, \hat{a}^{\ell-d}, s, s') &= P(a^\ell, c^\ell, s, s')P(b^\ell|a^\ell)1\{\hat{a}^{\ell-d} = h'(b^\ell, c^\ell, s')\} \\ &= P(a^\ell, c^\ell, s, s')[\prod_{i=1}^{\ell} P(b_i|a_i)]1\{\hat{a}^{\ell-d} = h'(b^\ell, c^\ell, s')\}, \end{aligned} \tag{17}$$

for every $b^\ell \in \mathcal{Y}^\ell$, and $\hat{a}^{\ell-d} \in \hat{\mathcal{X}}^{\ell-d}$. Let $(X^\ell, Y^\ell, U^\ell, \hat{X}^{\ell-d}, S, S')$ designate a random vector that is distributed according to this joint probability mass function. Now, accoding to the rate constraint:

$$\begin{aligned} nR &\geq \sum_{i=1}^{n} L(u_i) \\ &= \sum_{i=0}^{n/\ell-1} L(u_{i\ell+1}^{i\ell+\ell}) \\ &= \frac{n}{\ell} \sum_{s'\in\mathcal{S}'} \sum_{c^\ell \in \mathcal{C}^\ell(s')} P(c^\ell, s')L(c^\ell) \\ &\geq \frac{n}{\ell} \sum_{s'\in\mathcal{S}'} \sum_{c^\ell \in \mathcal{C}^\ell(s')} P(c^\ell, s') \log \frac{1}{P(c^\ell|s')} \end{aligned} \tag{18}$$

where the last step follows from the postulate that $\{L(c^\ell),\ c^\ell \in \mathcal{C}^\ell(s')\}$ satisfy the Kraft inequality for each s'. It follows then that $H(U^\ell|S') \leq \ell R$. Therefore,

$$\begin{aligned} \ell R &\geq H(U^\ell|S') \\ &= H(U^\ell, S') - I(S'; U^\ell) \\ &\geq H(U^\ell, S') - H(S') \\ &\geq H(U^\ell, S') - \log M. \end{aligned} \tag{19}$$

Next, observe that $u_{i\ell+1}^{i\ell+\ell}$ depends solely on $s_{i\ell+1}$ and $x_{i\ell+1}^{i\ell+\ell}$ but not on $y_{i\ell+1}^{i\ell+\ell}$, and so, $U^\ell \to (X^\ell, S) \to Y^\ell$ is a Markov chain. But since $S \to X^\ell \to Y^\ell$ is also a Markov chain (due to the DMC), then so is $(U^\ell, S) \to X^\ell \to Y^\ell$. In addition, $\hat{X}^{\ell-d} = h'(Y^\ell, U^\ell, S')$. It therefore follows that

$$\begin{aligned} \frac{1}{n}E\rho(x^n, \hat{X}^n) &\geq \frac{1}{\ell}E\rho(X^{\ell-d}, \hat{X}^{\ell-d}) \\ &\geq \frac{1}{\ell}[E\rho(X^\ell, \hat{X}^\ell) - d \cdot \rho_{\max}] \end{aligned} \tag{20}$$

where $\hat{X}^\ell$ is defined by concatenating $\hat{X}^{\ell-d}$ with a random d–vector in $\hat{\mathcal{X}}^d$ that is an arbitrary function of (Y^ℓ, U^ℓ, S'). Next, observe that

$$\begin{aligned} \ell R &\geq H(U^\ell, S') - \log M \\ &\geq H(U^\ell, S, S') - 2\log M, \end{aligned} \tag{21}$$

that is,

$$H(U^\ell, S, S') \leq \ell \left(R + \frac{2\log M}{\ell} \right), \tag{22}$$

and the Markovity of the chain $(U^\ell, S) \to X^\ell \to Y^\ell$ implies Markovity of $(U^\ell, S, S') \to X^\ell \to Y^\ell$. Moreover, for the purpose of deriving a lower bound, let us also allow h to depend on S too, i.e., $\hat{X}^\ell = h(Y^\ell, U^\ell, S, S')$. We now have the following lower bound to $\Delta_{M,d}(x^n, R)$:

$$\Delta_{M,d}(x^n, R) \geq \min_{U^\ell, S, S' h} \frac{1}{\ell} E\rho(X^\ell, h(Y^\ell, U^\ell, S, S')) - \frac{d\rho_{\max}}{\ell} \tag{23}$$

where the minimum is subject to the constraints:

$$H(U^\ell, S, S') \leq \ell \left(R + \frac{2\log M}{\ell} \right) \tag{24}$$

and

$$(U^\ell, S, S') \to X^\ell \to Y^\ell \text{ is a Markov chain.} \tag{25}$$

Now, observe that in this minimization problem U^ℓ, S, and S' appear always together. Let us define then $U \stackrel{\Delta}{=} (U^\ell, S, S')$ and further reduce this expression by taking the minimum of the distortion over all (U, h) subject to the constraints $H(U) \leq \ell(R + 2\log M/\ell)$, and $U \to X^\ell \to Y^\ell$, which is $\Delta_{X^\ell|Y^\ell}(R + 2\log M/\ell)$ by definition. This completes the proof of the Theorem 1. □

We now describe a (universal) block coding scheme that asymptotically (for large ℓ) achieves $\Delta_{\boldsymbol{X}|\boldsymbol{Y}}(R)$ and hence also $\Delta^*(\boldsymbol{x}, R)$ (for almost all values of R): Given x^n, compute its ℓ–th order empirical distribution, $\{P(a^\ell),\ a^\ell \in \mathcal{X}^\ell\}$ (which is $P(X^\ell)$), and find the RV U and the function h that achieve $\Delta_{X^\ell|Y^\ell}(R)$. The (stochastic) encoder applies the channel $P(U|X^\ell)$ to every ℓ–vector $x_{i\ell+1}^{i\ell+\ell}$ and then performs entropy coding according to the marginal of U, after transmitting a header that describes the entropy coding rule of the optimum U (which depends only on the marginal of U) and the function h, which together require no more than $\lceil \log(n/\ell + 1)^{\alpha^\ell} \rceil$ bits, which is log of the number of different empirical

distributions of superletters formed by ℓ–vectors). The decoder first decodes the header, then U, and finally, reconstructs the source by applying $\hat{X}^\ell = h(Y^\ell, U)$. The rate is upper bounded by

$$\frac{1}{n}\{\lceil \log[(k+1)^{\alpha^\ell}]\rceil + \frac{n}{\ell}[H(U)+1]\} \leq R + \frac{1}{\ell} + \frac{\alpha^\ell}{n}\log\left(\frac{n}{\ell}+1\right) \tag{26}$$

where we have used the fact that $H(U) \leq \ell R$. Clearly, if ℓ is large and $n >> \alpha^\ell$, this is arbitrarily close to R. The distortion $\Delta_{X^\ell|Y^\ell}(R)$ is maintained by definition.

Note that complexity of this scheme is mostly in the optimization over h and U, which is not negligible, but is a function of ℓ only. This is in contrast to the schemes proposed in [12],[13], which require an exhaustive search over sets of sequences of length $n(>> \alpha^\ell)$, namely, complexity that grows exponentially with n. The stochastic encoder $X^\ell \to U$ can also be implemented deterministically, but then the encoding complexity will be exponential in n: Select independently at random a set of $M = 2^{(n/\ell)[I(X^\ell;U)+\epsilon]}$ vectors $U^n(i)$, $i = 1, \ldots, M$. Given x^n, find a jointly typical vector $U^n(i)$ (in the superalphabet of ℓ–vectors), and transmit it using $(n/\ell)[I(X^\ell;U)+\epsilon] \leq (n/\ell)[H(U)+\epsilon]$ bits plus a (relatively small) header that describes the type class of x^n, from which the decoder can also figure out h on its own. By the Markov lemma, with high probability, (U, X^ℓ, Y^ℓ) will be jointly typical and hence $\hat{X}^n$ will satisfy the distortion constraint.

Discussion. Three comments are in order:

1. When $\boldsymbol{x}$ emerges from a discrete memoryless source (DMS), rather than being an individual sequence, it is well known that the distortion–rate function is the classical W–Z distortion rate function for that DMS, $\Delta^{WZ}_{X|Y}(R)$. Clearly, by analyzing the performance of block codes (rather than FSM's) on the given DMS, using the same tecnique as above, one can show that

$$\Delta^{WZ}_{X|Y}(R) = \inf_{\ell \geq 1} \Delta_{X^\ell|Y^\ell}(R), \tag{27}$$

   where now X^ℓ designates a random ℓ–vector from the source. This is true because for both sides of this equality, we have a direct theorem and a converse theorem.

2. In view of item no. 1 above and our direct thoerem, we have actually shown that it is possible to approach the W–Z rate–distortion function (of both a DMS and an individual sequence) without binning.

3. The above result extends to a model of scalable coding (successive refinement), in analogy to [8]. The details and the discussion appear in the Appendix.

4 Universalilty and the Critical Growth Rate of M

In the previous section, we have considered a reference class of encoders and decoders with a *fixed* number of states, M, and a *fixed* delay, d, and we have shown that the (operative) rate–distortion function w.r.t. this class can be approached by using (sufficiently long) block codes. In this section, we address a somewhat different question, pertaining to a regime that allows both the number of states and the delay to *grow* with n, the length of the input sequence x^n. That is, $M = M_n$ and $d = d_n$. For the sake of simplicity, we will be 'generous' with regard to the delay, allowing it to be maximum, namely, $d_n = d_{M_n}$, and then we focus on the following question: What is the highest growth rate of M_n as a function of n, below which it is still possible to universally attain (in a sense to be defined soon), using any general block code for n–sequences, the performance of the best encoder–decoder with M_n states and delay d_{M_n}?

For the sake of convenience, in this section, instead of confining ourselves to either rate–distortion functions or distortion–rate functions, we will treat rate and distortion in a more symmetric fashion by defining achievable pairs (R, Δ) in the spirit of the definitions in Section 1: Given x^n, a pair (R, Δ) is said to be M_n–*achievable* if there exists an M_n–state encoder $\mathcal{E} = (\mathcal{S}, f, g)$ and a M_n–state decoder $\mathcal{D} = (\mathcal{S}', f', g')$, with overall delay not exceeding $d_n = d_{M_n}$, such that $\sum_{i=1}^n L(u_i) \le nR$ and $\sum_{i=1}^n E\rho(x_i, \hat{X}_i) \le n\Delta$. Referring to the previous definitions, given x^n, the pair $(R, \Delta_{M_n, d_{M_n}}(x^n, R))$ is always M_n–achievable.

While the definition of an achievable pair (R, Δ) allows the choice of an encoder–decoder that depends on x^n, we now define the notion of a *universally* achievable pair (R, Δ): A pair (R, Δ) is said to be *universally achievable* w.r.t. $\{M_n\}_{n\ge 1}$ if for every $\epsilon > 0$, $\delta > 0$, and n sufficiently large, there exists an encoder–decoder that achieves rate less than or equal to $R + \epsilon$ and distortion less than or equal to $\Delta + \delta$ for every x^n for which (R, Δ) is M_n–achievable.

Let us assume, from now on, that M_n grows asymptotically linearly with some power of n, that is, $\lim_{n\to\infty}(\log M_n)/\log n = \theta$, where θ is a certain positive real. In this section, we are interested in the critical value of θ below which universal achievability of *every* pair (R, Δ) is guaranteed, but above this critical value, there exist pairs (R, Δ) that are not

universally achievable.

The following two theorems tell us that this critical value is $\theta = 1$, in other words, the critical asymptotic growth rate of M_n is *linear*.

Theorem 2 *(Direct Theorem): If $\theta < 1$, then every (R, Δ) is universally achievable.*

Proof. Assume, for the sake of simplicty and without essential loss in generality, that $M_n = n^\theta$, $\theta < 1$, and consider the following mechanism for an encoding x^n: The encoder examines all possible pairs $\{(\mathcal{E}, \mathcal{D})\}$ with M_n states, on the given x^n, and computes the coding rate and the expected distortion (w.r.t. the randomness of the known channel from x^n to y^n). If it finds an encoder–decoder pair $(\mathcal{E}^*, \mathcal{D}^*)$ that achieves a specified rate–distortion pair (R, Δ), it first transmits a header with the description of $\mathcal{D}^*$ and then encodes x^n using $\mathcal{E}^*$ (if it does find such an encoder then (R, Δ) is not M_n–achievable in the first place). The decoder, after decoding the index of the decoder $\mathcal{D}^*$, uses this decoder to produce the reproduction $\hat{x}^n$ based on y^n and the remaining part of the bitstream. Obviously, the distortion associated with such an encoder is the same as that of $(\mathcal{E}^*, \mathcal{D}^*)$, namely, less than or equal to Δ. The rate is the same as that of $(\mathcal{E}^*, \mathcal{D}^*)$ (which means less than or equal to R) plus the rate associated with the header. Thus, it remains to show that the normalized redundancy associated with the header goes to zero as $n \to \infty$ whenever $\theta < 1$. To this end, we now evaluate the number of bits necessary to describe the decoder with $\mathcal{D}^*$.

First, observe that for each $s'_i \in \mathcal{S}'$, u_i may take values in a prefix code $\mathcal{C}(s'_i)$, which can be described by a tree. As each such tree contains at most α leaves, and there are $(k-1)!$ different trees with k leaves, the total number of possible trees is $K = \sum_{k=1}^{\alpha}(k-1)!$, thus the description of each such tree takes $\lceil \log K \rceil$ bits, and since such a tree should be specified for every $s' \in \mathcal{S}'$ this takes $M_n \cdot \lceil \log K \rceil$ bits altogether. Next, the function f' should be described. As there are no more than $\gamma^{M_n \alpha\beta}$ functions $\{f'\}$ for every possible tree, the description of f' takes $\lceil M_n \alpha\beta \log \gamma \rceil$ bits. Similarly, the description of g' requires at most $\lceil M_n \alpha\beta \log M_n \rceil$ bits. Thus, the total number of bits associated with the header is then $O(n^\theta \log n)$, which when normalized by n (to get the redundancy per source symbol), tends to zero as $n \to \infty$, since θ is assumed strictly less than unity. This completes the proof of the Theorem. $\square$

For the converse part, we will make the additional assumptions that $\hat{\mathcal{X}} = \mathcal{X}$, $\mathcal{X}$ is

a group with an addition operation (modulo α), and that the distortion function $\rho(x,\hat{x})$ depends on x and $\hat{x}$ only via their difference $x-\hat{x}$ (w.r.t. the group arithmetic). We will then denote $\rho_0(x-\hat{x})=\rho(x,\hat{x})$, where $\rho_0:\mathcal{X}\to\mathbb{R}^+$.

Theorem 3 *(Converse Theorem): Under the assumptions of the last paragraph, if $\theta>1$ there exist pairs (R,Δ) that are not universally achievable.*

Discussion. An alternative question with regard to universality w.r.t. FSM's with a growing number of states, which is closer in spirit to the results of Section 3, could have been the following: What is the critical growth rate of M_n that still allows $\Delta_{M_n,d_{M_n}}(x^n,R)$ to be achievable by a universal code for all x^n? This definition is seemingly stronger because it appears to give rise to 'adaptation' of the distortion to the given x^n rather than making a commitment to a fixed distortion level Δ, regardless of x^n. However, it is easy to see that both our direct theorem and converse theorem are suitable for this definition too, and therefore, so is the conclusion regarding the linear critical rate of M_n. The direct part would be the same, but with $(\mathcal{E}^*,\mathcal{D}^*)$ being defined as the encoder–decoder pair that achieves $\Delta_{M_n,d_{M_n}}(x^n,R)$, or alternatively, $R_{M_n,d_{M_n}}(x^n,\Delta)$, the M_n–state rate–distortion function, defined in a dual manner. Regarding the converse, as we demonstrate in the proof of Theorem 3 below, for every given $\theta>1$, there is a pair (R,Δ) which is M_n–achievable for certain sequences, and hence $\Delta_{M_n,d_{M_n}}(x^n,R)\le\Delta$, or, equivalently, $R_{M_n,d_{M_n}}(x^n,\Delta)\le R$. But no single encoder performs even close to $R_{M_n,d_{M_n}}(x^n,\Delta)$ simultaneously for all these sequences.

Proof of Theorem 3. Assume, again, that $M_n=n^\theta$. We will now show that for $\theta>1$, there exists a rate–distortion pair (R,Δ) which is not universally achievable. For a given Δ, let

$$\phi(\Delta)=\max\{H(Z):\ E\rho_0(Z)\le\Delta\},\tag{28}$$

where $H(Z)$ is the entropy of an RV Z taking on values in $\mathcal{X}$. Let $R_{X|Y}^{WZ}(\Delta)$ denote the W–Z rate–distortion function of the memoryless uniform source X with side information Y (generated by the given DMC $P(y|x)$) w.r.t. ρ. For a given Δ, and a given $\theta>1$, let

$$R=\frac{\phi(\Delta)}{\theta-1}\tag{29}$$

and select Δ to be sufficiently small such that

$$R=\frac{\phi(\Delta)}{\theta-1}<R_{X|Y}^{WZ}(\Delta)\le R_X(\Delta)=\log\alpha-\phi(\Delta),\tag{30}$$

where $R_X(\Delta)$ is the ordinary rate–distortion function (without side information) of the memoryless uniform source X w.r.t. ρ. We wish to show that for such a choice of R and Δ, there exists a set of sequences $\{x^n\}$ for each of which (R, Δ) is M_n–achievable, but on the other hand, there is no *single* code that simultaneously achieves (R, Δ) for all sequences in this set.

For the above choice of R and Δ, consider a random process defined as follows. Let m be the the solution to the equation

$$2^{Rm} = \frac{n}{m}, \tag{31}$$

and assume that this solution is integer. Further, let $\mathcal{F}$ be a set of 2^{mR} m–vectors $\mathcal{F} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{2^{mR}}\}$, $\boldsymbol{u}_i, \in \mathcal{X}^m$, $i = 1, \ldots, 2^{mR}$. Assuming that m divides n (i.e., 2^{mR} is integer), let x^n be formed by concatenatingg n/m m–vectors $\{x^m(i),\ i = 1, \ldots, n/m\}$, where

$$x^m(i) = u^m(i) + z^m(i), \quad i = 1, \ldots, n/m, \tag{32}$$

$u^m(i)$ is an arbitrary member of $\mathcal{F}$, and $z^m(i) \in \mathcal{X}^m$ is a vector of i.i.d. random variables, each component of which is distributed according to the distribution P^* on $\mathcal{X}$ which achieves $\phi(\Delta)$. Now, we further assume that $\mathcal{F}$ is a good[3] code for the additive memoryless channel $X = U + Z$, where U designates the channel input at rate R, $Z \sim P^*$ is the memoryless noise, and X stands for the channel output. Since R is assumed less than the capacity of this channel, given by $C = R_X(\Delta) = \log \alpha - \phi(D)$, then such a good code exists. This means that upon observing $x^m(i)$, one can identify $u^m(i)$ correctly with high probability, provided that m is large.

Consider next a *block* code of length m operating on x^n as follows: For every $i = 1, \ldots, n/m$, the encoder first decodes $u^m(i)$ from $x^m(i)$, and then transmits a description of the decoded version, say $\hat{u}^m(i)$, using $\log |\mathcal{F}| = mR$ bits. The decoder, in turn, reconstructs $\hat{x}^m(i) = \hat{u}^m(i)$ (without using the side information). Since $\hat{u}^m(i) = u^m(i)$ with high probability, then the distortion between $x^m(i)$ and $\hat{x}^m(i)$, is about Δ, because the noise $z^m(i)$ is distributed according to P^*, whose ρ_0–moment does not exceed Δ. This means that the pair (R, Δ) is essentially achievable by a (non–universal) block code of size m.

Now, since the set of typical input m–vectors to the block code is of size that is of the exponential order of $2^{m[R+\phi(\Delta)]}$, this block code can be implemented by an FSM with with essentially no more than $m2^{m[R+\phi(\Delta)]}$ states. This can be done by constructing an

[3]In the sense of small error probability.

(incomplete) α–ary context–tree with $2^{m[R+\phi(\Delta)]}$ leaves, corresponding to the various typical sequences, where the state set is the set of nodes plus the leaves of this incomplete tree.[4] Thus, the number of states as a function of n is given by

$$\begin{aligned} M_n &\leq m2^{m[R+\phi(\Delta)]} \\ &\leq [m2^{mR}]^{[1+\phi(\Delta)/R]} \\ &= n^{1+\phi(\Delta)/R} \\ &= n^{\theta}. \end{aligned} \tag{33}$$

This means that we have shown that for the above choice of R and Δ, the pair (R,Δ) is n^{θ}–achievable for every x^n that is typical to the above defined process.

We now argue that no block–code of length n can attain (R,Δ) simulatenously for *all* typical sequences of the process (32), and for all 'good' channel codes $\{\mathcal{F}\}$, which yield error probability less than $2^{-m[E_r(R)-\delta]}$, where $E_r(R)$ is the random coding error exponent [3] of the channel $X=U+Y$ w.r.t. the uniform random coding input distribution, and $\delta\in(0,E_r(R))$ is arbitrary.[5] Furthermore, we show that even $(R_{X|Y}^{WZ}(\Delta)-\epsilon,\Delta)$, for arbitrarily small (but fixed) $\epsilon>0$, cannot be simultaneously attained for all those sequences (recall that $R_{X|Y}^{WZ}(\Delta)>R$).

Let us denote the set of all the typical u–sequences by $\mathcal{G}$, i.e., $\mathcal{G}$ is the set of all sequences $\{u^n\}$ whose segments $\{u^m(i)\}$ form a code (for the channel $X=U+Y$) whose error probability, $P_e(u^n)$, is below $2^{-m[E_r(R)-\delta]}$, and observe that

$$\Pr\{\mathcal{G}^c\}=\Pr\{u^n:\ P_e(u^n)>2^{-m[E_r(R)-\delta]}\}\leq\frac{E\{P_e(U^n)\}}{2^{-m[E_r(R)-\delta]}}\leq\frac{2^{-mE_r(R)}\}}{2^{-m[E_r(R)-\delta]}}=2^{-m\delta}, \tag{34}$$

where the first inequality follows from the Chebychev inequality. Now, assume conversely, that there exists a source code that *does* achieve $(R_{X|Y}^{WZ}(\Delta)-\epsilon,\Delta)$ for all x^n induced by all $u^n\in\mathcal{G}$ and all typical z^n–sequences, namely, sequences for which (a very high fraction of) the m-segments $\{z^m(i)\}$ are P^*–typical. Then, we have

$$\frac{1}{|\mathcal{G}|}\sum_{u^n\in\mathcal{G}}EL(u^n+Z^n)\leq n[R_{X|Y}^{WZ}(\Delta)-\epsilon] \tag{35}$$

[4] To be precise, one should add m more states corresponding to a modulo–m time counter, in order to idle for non-typical sequences (which are not terminated by the leaves) until the end of the block and then submit an error message.

[5] Such a code may harm the rate (beyond R) and the distortion (beyond Δ) only for fraction $2^{-m[E_r(R)-\delta]}$ of the m–segments.

and

$$\frac{1}{|\mathcal{G}|}\sum_{u^n\in\mathcal{G}} E\rho(u^n+Z^n,\hat{X}^n)\le n\Delta, \tag{36}$$

and so, for every $\lambda\ge 0$,

$$\frac{1}{|\mathcal{G}|}\sum_{u^n\in\mathcal{G}} [EL(u^n+Z^n)+\lambda E\rho(u^n+Z^n,\hat{X}^n)]\le n[R_{X|Y}^{WZ}(\Delta)-\epsilon+\lambda\Delta], \tag{37}$$

where the inner expectations are w.r.t. the uniform distribution over all typical z–sequences. Consider now a *random* selection of the $2^{Rm}=n/m$ members of $\mathcal{F}$ independently and with uniform distribution over $\mathcal{X}^m$. On the one hand, this induces the uniform distribution over $\mathcal{X}^n$, for which we know, by the converse to the W–Z rate–distortion theorem, that either

$$\frac{1}{\alpha^n}\sum_{u^n} E\rho(u^n+Z^n,\hat{X}^n)\ge n\Delta, \tag{38}$$

or

$$\frac{1}{\alpha^n}\sum_{u^n} EL(u^n+Z^n)\ge nR_{X|Y}^{WZ}(\Delta), \tag{39}$$

where the inner expectations over Z^n are as before (note that as U^n is uniformly distributed then so is $X^n=U^n+Z^n$ regardless of Z^n, which is independent). It then follows that there exists $\lambda\ge 0$ (which is bounded independently of n) such that

$$\frac{1}{\alpha^n}\sum_{u^n}[EL(u^n+Z^n)+\lambda E\rho(u^n+Z^n,\hat{X}^n)]\ge n[R_{X|Y}^{WZ}(\Delta)+\lambda\Delta]. \tag{40}$$

To see why this is true, let us denote $n\Delta'\stackrel{\Delta}{=}\alpha^{-n}\sum_{u^n}E\rho(u^n+Z^n,\hat{X}^n)$, and then the left–hand side of eq. (40) is lower bounded by $n[R_{X|Y}^{WZ}(\Delta')+\lambda\Delta']$. Now, if $\Delta'\le\Delta$, then eq. (40) clearly holds for $\lambda=0$. Else, if $\Delta'>\Delta$, it obviously holds for

$$\begin{aligned}\lambda &= \frac{R_{X|Y}^{WZ}(\Delta)-R_{X|Y}^{WZ}(\Delta')}{\Delta'-\Delta}\\ &\le \frac{R_{X|Y}^{WZ}(0)-R_{X|Y}^{WZ}(\Delta)}{\Delta-0}\\ &= \frac{H(X|Y)-R_{X|Y}^{WZ}(\Delta)}{\Delta}\end{aligned} \tag{41}$$

where the inequality follows from the convexity of the Wyner–Ziv rate–distortion function [10], [1, p. 439, Lemma 14.9.1]. Therefore, in either case, the value of λ that satisfies eq. (40) is bounded independently of n (for $\Delta>0$). For this value of λ, we then have

$$\sum_{u^n\in\mathcal{G}^c}[EL(u^n+Z^n)+\lambda E\rho(u^n+Z^n,\hat{X}^n)]$$

$$
\begin{aligned}
&= \sum_{u^n}[EL(u^n+Z^n)+\lambda E\rho(u^n+Z^n,\hat{X}^n)] - \\
&\quad \sum_{u^n\in\mathcal{G}}[EL(u^n+Z^n)+\lambda E\rho(u^n+Z^n,\hat{X}^n)] \\
&\geq n\alpha^n[R^{WZ}_{X|Y}(\Delta)+\lambda\Delta] - n|\mathcal{G}|[R^{WZ}_{X|Y}(\Delta)-\epsilon+\lambda\Delta] \\
&\geq n\alpha^n[R^{WZ}_{X|Y}(\Delta)+\lambda\Delta] - n\alpha^n[R^{WZ}_{X|Y}(\Delta)-\epsilon+\lambda\Delta] \\
&= n\epsilon\alpha^n. \qquad (42)
\end{aligned}
$$

On the other hand, assuming that $\max_x \rho_0(x) = \rho_{\max} < \infty$, and that $\max_{x^n} L(x^n) \leq nL$ for some finite $L > 0$ (otherwise, long codewords would be better transmitted uncompressed plus an extra bit to tell that they are uncompressed), then we have

$$
\begin{aligned}
\sum_{u^n\in\mathcal{G}^c}[EL(u^n+Z^n)+\lambda E\rho(u^n+Z^n,\hat{X}^n)] &\leq n|\mathcal{G}^c|(L+\lambda\rho_{\max}) \\
&= \Pr\{\mathcal{G}^c\}n\alpha^n(L+\lambda\rho_{\max}) \\
&\leq n2^{-m\delta}\alpha^n(L+\lambda\rho_{\max}). \qquad (43)
\end{aligned}
$$

Comparing now the right–most sides of eqs. (42) and (43), we get $\epsilon \leq (L+\lambda\rho_{\max})2^{-m\delta}$, which is a contradiction for all large n (and m) since $\epsilon > 0$ was assumed a constant (and λ is bounded). Thus, we have disproved the existence of a code that achieves (R,Δ) simulateneously for all typical sequences of the process described above. This completes the proof.

Appendix

Extending the Results of Section 3 to Successive Refinement Codes

Consider a two–stage coding scheme with a successive refinement structure. The first stage is as in Section 3. In the second stage, there is an additional finite–state encoder $\mathcal{E}'$ that transmits a refining description $\boldsymbol{v} = (v_1, v_2, \ldots)$ of variable–length binary strings (similarly as $\boldsymbol{u}$), and an additional finite–state decoder $\mathcal{D}'$ that has access to both $\boldsymbol{u}$ and $\boldsymbol{v}$ as well as to another SI sequence $\boldsymbol{z} = (z_1, z_2, \ldots)$. It is assumed that

$$
P(y^n, z^n|x^n) = \prod_{i=1}^{n} P(y_i, z_i|x_i), \quad n = 1, 2, \ldots \qquad (A.1)
$$

More precisely, the description of the second stage is as follows: When the sequence $\boldsymbol{x}$ is sequentially fed into a variable–rate finite–state encoder $\mathcal{E}' = (\mathcal{T}, p, g)$, this encoder

generates an infinite sequence of binary strings of variable length, $\boldsymbol{v} = (v_1, v_2, \ldots)$, while going through an infinite sequence of states $t_1, t_2, \ldots$ according to

$$\begin{aligned} v_i &= p(t_i, x_i), \\ t_{i+1} &= q(t_i, x_i), \quad i = 1, 2, \ldots \end{aligned} \tag{A.2}$$

where the initial state t_1 is assumed to be a certain fixed member of $\mathcal{T}$. At the same time and in a similar manner, the second–stage finite–state decoder $\mathcal{D}' = (\mathcal{T}', p', q')$ sequentially maps $\boldsymbol{u}$, $\boldsymbol{v}$ and $\boldsymbol{z}$ to an infinite reproduction sequence $\tilde{x}_1, \tilde{x}_2, \ldots$, using the recursion

$$\begin{aligned} \tilde{x}_{i-d'} &= p'(t'_i, u_i, v_i, z_i), \quad i = d'+1, d'+2, \ldots \\ t'_{i+1} &= q'(t'_i, u_i, v_i, z_i), \quad i = 1, 2, \ldots. \end{aligned} \tag{A.3}$$

where d' (positive integer) is the second–stage encoding–decoding delay, and the initial state t'_1 is assumed to be a certain fixed member of $\mathcal{T}'$. Similarly to the model of variable–rate coding of the first stage, it is assumed that at each time instant i, when state decoder is at state t'_i and reads[6] the current first–stage codeword u_i, it is able to isolate the current input codeword v_i from the following codewords of the compressed bitstream, $v_{i+1}, v_{i+2}, \ldots$. To this end, we allow a prefix code $\mathcal{C}'(t', u)$ associated with each $t' \in \mathcal{T}'$ and $u \in \mathcal{C}(s')$. For every $v \in \mathcal{C}'(t')$, let $L'(v)$ denote the length of v (in bits). We then assume that the Kraft inequality

$$\sum_{v \in \mathcal{C}'(t', u)} 2^{-L'(v)} \leq 1 \tag{A.4}$$

holds for all $t' \in \mathcal{T}'$, $u \in \mathcal{C}(s')$. Again, this discussion continues to apply when single codewords are replaced by ℓ–vectors, v^ℓ, formed by concatenating ℓ (legitimate) codewords successively. In this case, let $[\mathcal{C}']^\ell(t', u^l)$ denote the supercode formed by all $\{v^\ell\}$ that originate from (t', u^l). Clearly, since the components of v^ℓ can be identified recursively, the supercode satisfies the Kraft inequality as well w.r.t. the length function $L'(v^\ell) = \sum_{i=1}^{\ell} L'(v_i)$.

Let $\Delta'(x^n, \mathcal{E}, \mathcal{D})$ denote the expected distortion $\frac{1}{n} E\rho'(x^n, \tilde{X}^n) \stackrel{\Delta}{=} \frac{1}{n} \sum_{i=1}^{n} E\rho'(x_i, \tilde{X}_i)$ associated with the second–stage encoding and decoding of x^n by $(\mathcal{E}, \mathcal{D})$ and $(\mathcal{E}', \mathcal{D}')$, where the expectation is w.r.t. the DMC's. For a given x^n and rate pair $(R, \Delta R)$, a distortion pair (D_1, D_2) is said to be (M, d, d')–achievable if there exist encoder–decoders $(\mathcal{E}, \mathcal{D})$ and

[6]The second stage decoder can keep a copy of s''_i of footnote no. 2 as it has access to $\{u_i\}$ as well.

$(\mathcal{E}', \mathcal{D}')$, having no more than M states each, with first–stage delay not exceeding d, second–stage delay not exceeding d', and which satisfy the rate constraints:

$$\begin{aligned} \sum_{i=1}^{n} L(u_i) &\leq nR \\ \sum_{i=1}^{n} L'(v_i) &\leq n\Delta R \end{aligned} \tag{A.5}$$

and

$$\begin{aligned} \Delta(x^n, \mathcal{E}, \mathcal{D}) &\leq D_1 \\ \Delta'(x^n, \mathcal{E}', \mathcal{D}') &\leq D_2 \end{aligned} \tag{A.6}$$

Let $\Delta_{M,d,d'}(x^n, R, \Delta R)$ denote the set of distortion pairs (D_1, D_2) that are (M, d, d')–achievable for x^n.

While the definition of $\Delta_{M,d,d'}(x^n, R, \Delta R)$ is operational, we next define an informational achievable region $\Delta_{X^\ell|Y^\ell,Z^\ell}(R, \Delta R)$ (with X^ℓ being distributed according to the empirical distirbution of ℓ–vectors) as follows: $(D_1, D_2) \in \Delta_{X^\ell|Y^\ell,Z^\ell}(R, \Delta R)$ iff there exist random variables U, V that satisfy the Markov relation $(U, V) \to X^\ell \to (Y^\ell, Z^\ell)$ and functions h and h' such that

$$\begin{aligned} E\rho(X^\ell, h(Y^\ell, U)) &\leq \ell D_1 \\ E\rho'(X^\ell, h'(Z^\ell, U, V)) &\leq \ell D_2 \\ H(U) &\leq \ell R \\ H(V|U) &\leq \ell \Delta R \end{aligned} \tag{A.7}$$

In the theorem below, which is an extension of Theorem 1, $\Delta_{X^\ell|Y^\ell,Z^\ell}(R, \Delta R) - (\delta, \delta')$ means the set $\{(D_1 - \delta, D_2 - \delta') : \ (D_1, D_2) \in \Delta_{X^\ell|Y^\ell,Z^\ell}(R, \Delta R)\}$.

Theorem 4 *: For every positive integers n and ℓ such that ℓ divides n:*

$$\Delta_{M,d,d'}(x^n, R, \Delta R) \subseteq \Delta_{X^\ell|Y^\ell,Z^\ell}\left(R + \frac{2\log M}{\ell}, \Delta R + \frac{\log M}{\ell}\right) - \frac{1}{\ell}(\rho_{\max} d, \rho'_{\max} d'). \tag{A.8}$$

where $\rho'_{\max} \stackrel{\Delta}{=} \max_{x,\tilde{x}} \rho'(x, \tilde{x})$.

Proof. The first stage is as before. As for the second stage, similarly to (18), (19) and (21), we have

$$\ell \Delta R \ \geq \ H(V^\ell | U^\ell, T')$$

$$\begin{aligned} &\geq H(V^\ell, T'|U^\ell, S, S', T') \\ &\geq H(V^\ell, T, T'|U^\ell, S, S') - \log M \end{aligned} \tag{A.9}$$

Also, $(U^\ell, V^\ell) \to (X^\ell, S, S', T, T') \to (Y^\ell, Z^\ell)$ is a Markov chain. Again, since $(S, S', T, T') \to X^\ell \to (Y^\ell, Z^\ell)$ is also a Markov chain, then so is $(U^\ell, S, S', V^\ell, T, T') \to X^\ell \to Y^\ell \to Z^\ell$. The reconstructed output $\tilde{X}^{\ell-d'}$ is a function of $(U^\ell, V^\ell, T', Z^\ell)$ which is a special case of a function of $(U^\ell, S, S', V^\ell, T, T', Z^\ell)$ and the distortion at the second stage is then lower bounded by $\frac{1}{\ell}E\rho'(X^\ell, \tilde{X}^\ell) - \rho'_{\max}d'/\ell$ as before. Hence, defining $U = (U^\ell, S, S')$ and $V = (V^\ell, T, T')$, we have found RV's (U, V) and functions h and h' that satisfy the conditions for $(D_1 - \rho_{\max}d/\ell, D_2 - \rho'_{\max}d'/\ell)$ being in $\Delta_{X^\ell|Y^\ell,Z^\ell}(R + 2\log M/\ell, \Delta R + 2\log M/\ell)$ provided that $(D_1, D_2) \in \Delta_{M,d,d'}(x^n, R, \Delta R)$. $\square$

The achievability is conceptually simple. Again, the first stage is as before. The second stage is a conditional version of the first where both encoder and decoder have access to the already decoded U.

Finally, a few comments are in order, in addition to the comments made at the Discussion of Section 3:

1. While in [8] there is no apparent way to generalize the results to more than two stages (unless the SI's are identical), here the extension is straightforward.

2. Unlike in [8], here there is no need for the Markov structure $X \to Z \to Y$.

3. The alphabet sizes required for U and V are $|\mathcal{U}| \leq \alpha^\ell + 3$ and $|\mathcal{V}| \leq \alpha^\ell \cdot |\mathcal{U}| + 1$.

Acknowledgement

Neri Merhav would like to thank Tsachy Weissman for useful discussions at the early stages of this work.

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991.

[2] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inform. Theory*, vol. IT–37, no. 2, pp. 269–275, March 1991.

[3] R. G. Gallager, *Information Theory and Reliable Communication*, J. Wiley & Sons, 1968.

[4] A. Lempel and J. Ziv, "Compression of two–dimensional data," *IEEE Trans. Inform. Theory*, vol. IT–32, no. 1, pp. 2–8, January 1986.

[5] B. Rimoldi, "Successive refinement of information: characterization of achievable rates," *IEEE Trans. Inform. Theory*, vol. IT–40, no. 1, pp. 253–259, January 1994.

[6] E. Ordentlich, T. Weissman, M. J. Weinberger, A. Somekh–Baruch, and N. Merhav, "Discrete universal filtering through incremental parsing," *Proc. 2004 Data Compression Conference (DCC '04)*, pp. 352–361, Snowbird, UT, March 2004.

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

[8] Y. Steinberg and N. Merhav, "On successive refinement for the Wyner–Ziv problem," *IEEE Trans. Inform. Theory*, vol. 50, no. 8, pp. 1636–1654, August 2004.

[9] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. J. Weinberger, "Universal discrete denoising: known channel," *IEEE Trans. Inform. Theory*, vol. 51, no. 1, pp. 5–28, January 2005.

[10] A. D. Wyner and J. Ziv, "The rate–distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. IT–22, no. 1, pp. 1–10, January 1976.

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

[12] J. Ziv, "Distortion–rate theory for individual sequences," *IEEE Trans. Inform. Theory*, vol. IT–26, no. 2, pp. 137–143, March 1980.

[13] 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.

[14] 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.