Finite–State Source–Channel Coding for Individual Source Sequences with Source Side Information at the Decoder

Neri Merhav

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

Abstract

We study the following semi–deterministic setting of the joint source–channel coding problem: a deterministic source sequence (a.k.a. individual sequence) is transmitted via a memoryless channel, using delay-limited encoder and decoder, which are both implementable by periodically–varying finite-state machines, and the decoder is granted with access to side information, which is a noisy version of the source sequence. We first derive a lower bound on the achievable expected distortion in terms of the empirical statistics of the source sequence, the number of states of the encoder, the number of states of the decoder, their period, and the overall delay. The bound is shown to be asymptotically achievable by universal block codes in the limit of long blocks. We also derive a lower bound to the best achievable excess–distortion probability and discuss situations where it is achievable. Here, of course, source coding and channel coding cannot be completely separated without loss of optimality. Finally, we outline a few extensions of the model considered, such as: (i) incorporating a common reconstruction constraint, (ii) availability of side information at both ends, and (iii) extension to the Shannon channel with causal state information at the encoder. This work both extends and improves on earlier work of the same flavor (Ziv 1980, Merhav 2014), which focused only on the expected distortion, without side information at either end, and without the above mentioned additional ingredients.

Index Terms: Wyner–Ziv problem, Shannon channel, causal state information, individual sequences, separation theorem, joint source–channel coding, finite–state machine, delay, excess–distortion exponent.

1 Introduction

In a collection of works that appeared during the late seventies and eighties of the previous century, Ziv [26], [27], [28], and Ziv and Lempel [13], [30], have established a fascinating theory of universal source coding for deterministic sequences (a.k.a. individual sequences) by means of encoders/decoders that are implementable using finite–state machines. Specifically, in [26] Ziv addressed the issue of fixed–rate, universal (nearly) lossless compression of deterministic source sequences using finite–state encoders and decoders, which was later further developed to the celebrated Lempel–Ziv algorithm [13], [30]. In [27], the model setting of [26] was broadened to lossy transmission over both clean and noisy channels (subsections II.A and II.B therein, respectively), where in the noisy case, the channel was modeled as an ordinary, probabilistic memoryless channel, as opposed to the source sequence, that was still assumed deterministic. Henceforth, we will refer to this type of setting as a *semi–deterministic* setting, similarly as in [16].

Subsequently, the results of the first part of [27] (clean channels) were further elaborated in other directions, such as exploiting side information in a scenario of almost lossless source coding, where the side information is modeled too as being deterministic [28], i.e., a deterministic analogue of Slepian–Wolf coding was investigated in [28]. More than two decades later, this setup was extended to the lossy case [19], that is, a semi–deterministic counterpart of Wyner–Ziv coding, where the source to be compressed is deterministic, but the side information available to the decoder is generated from the source sequence via a discrete memoryless channel (DMC). The model of [19] was still a pure source–coding model, where the main channel was clean, and the encoding model allowed variable–length coding. In [16], a few inaccuracies in the coding theorem for noisy channels in [27, Subsection II.B] were corrected, and it was strengthened and refined from several aspects. Among other things, in [16], only the decoder was assumed to be a finite–state machine, while the encoder was allowed to be rather general (as opposed to [27], where the encoder was assumed to be a finite–state machine too). Also, the finite–state decoder of [16] was allowed to be periodically time–varying with a given period length, ℓ, along with a modulo–ℓ time counter (clock).

In this work, we further develop the findings of [16] and [19] in a few directions, and at the same time, we also take the opportunity to correct some (minor) imprecisions in [16] and improve the rigor of the derivation, as well as the tightness of the converse bound. In our present model, we are back to assume that both the encoder and the decoder are finite–state machines (similarly as in [27] and [28]), but as in [16], we allow them to be periodically time–varying (with the same period), having a limited delay, and we also allow the decoder to access side information, which is a noisy version (corrupted by a DMC) of the deterministic source sequence to be conveyed – see Fig. 1. In other words, it is a semi–deterministic setting of a joint source–channel coding problem that combines the semi–deterministic version of the Wyner–Ziv (W-Z) source model with the DMC, in analogy to the purely stochastic version of this model [20]. The W-Z channel can be motivated by the uncoded transmission of the systematic part of a systematic code (see also [17], [20]).

At first glance, one might wonder about this asymmetric modeling approach of the semi-deterministic setting (both here and in the earlier works, [16], [19], [27]), where the source is regarded completely deterministically, without any statistical assumptions, while the channels (namely, both the main channel and the Wyner–Ziv channel) are modeled probabilistically, exactly as in the classical tradition of the information theory literature. The motivation for this distinction, is that in many frequently encountered situations, the channels obey some relatively well–understood physical laws that govern the underlying noise processes, which can be reasonably well be modeled probabilistically, whereas the source to be conveyed is very different in nature. Indeed, in many applications, the source is a man–made data file (or a group of files), generated using artificial means. These include computer–generated images and video streams, texts of various types, audio signals (such as music), sequences of output results from computer calculations, and any combinations of those. It is simply inconceivable to use ordinary probabilistic models for such sources.

For the above described semi-deterministic model, we first derive a lower bound to the minimum achievable expected distortion between the source and its reconstruction at the decoder. Our main result is a lower bound to the best achievable distortion, which depends on: (i) the given source sequence, (ii) the capacity of the main channel, (iii), the W-Z channel, (iv) the period ℓ, (v) the number of states of the encoder, s_{e}, and (vi) the number of states of the decoder, s_{d}. It turns out that the lower bound depends on s_{e} *very differently* than on s_{d}: The dependence on s_{e} is much weaker and it completely vanishes as the length n of the source sequence tends to infinity. In other words, as far as the lower bound is concerned, s_{d} is a much more significant resource than s_{e}. This asymmetric behavior of the lower bound is interesting and not trivial. The bound is also tighter than in [27]. It can be universally asymptotically achieved by separate W-Z source coding and channel coding, using long block codes, provided that both d and $\log s_{\rm d}$ are small compared to ℓ.

In addition to the the expected distortion, we also address a related, but different figure of merit – the probability of excess distortion, similarly as in [5] and [14]. We first derive a lower bound to this probability for the simpler model setting without side information, as in [16] and [27]. We relate it to the probability of excess distortion in the purely probabilistic setting [4], [5], and then discuss when this bound is asymptotically achievable. Subsequently, we extend the scope to the above–described model with decoder side information. Achievability is discussed too, but it should be pointed out that our emphasis, in this work, is on fundamental limits and lower bounds, more than on achievability.

In the last part of this work, we discuss a few variants of the model, where more explicit results can be stated, such as the case where the source side information is available at the encoder too. The case where a common reconstruction constraint is imposed, following [23], is also presented. Finally, we discuss an extension where the ordinary DMC is replaced by the Shannon channel with causal state information at the encoder [21].

The outline of the remaining part of this article is as follows. In Section 2, we establish notation conventions and formalize the problem setting and the objectives. In Section 3, we provide a few additional definitions in order to establish the preparatory background needed to state the main results, and we also provide a preliminary result, for the case of an ordinary DMC and without side information. In Section 4, we provide the extension that incorporates source side information, and finally, in Section 5, we outline a few modifications and extensions of our setting as described in the previous paragraph.

2 Notation Conventions and Problem Formulation

2.1 Notation Conventions

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. 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 n–th order Cartesian power of $\mathcal{Y}$, which is the alphabet of each component of this vector. An infinite sequence will be denoted by the bold face font, for example, $\boldsymbol{u} = (u_1, u_2, \ldots)$. The notation $\boldsymbol{u}_i$, on the other hand, will be used to denote the i–th ℓ–block $(u_{i\ell+1}, u_{i\ell+2}, \ldots, u_{i\ell+\ell})$. 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. If, in addition, $j = 1$, the superscript will be omitted too, and the notation will be simply x.

Owing to the semi-deterministic modeling approach, we distinguish between two kinds of random variables: ordinary random variables (or vectors), governed by certain given probability distributions (like the channel output vector), and auxiliary random variables that emerge from empirical distributions associated with certain sequences. Random variables of the second kind will be denoted using 'hats'. For example, consider a deterministic sequence $u^n = (u_1, \ldots, u_n)$. Then, $\hat{U}^\ell = (\hat{U}_1, \ldots, \hat{U}_\ell)$ designates an auxiliary random vector, 'governed' by the empirical distribution extracted from the non-overlapping ℓ–blocks of u^n, provided that ℓ divides n (this empirical distribution will be defined precisely in the sequel). We denote this empirical distribution by $P_{\hat{U}^\ell} = \{P_{\hat{U}^\ell}(u^\ell),\ u^\ell \in \mathcal{U}^\ell\}$. The use of empirical distributions, however, will not be limited to deterministic sequences only. It could be defined also for realizations of random sequences. For example, if Y^n is a random sequence, $\hat{Y}^\ell$ would designate the auxiliary random ℓ-vector associated with a given realization, y^n, of Y^n. In this case, the empirical distribution, $P_{\hat{Y}^\ell}$, is of course, itself random, but it may converge to the true ℓ–th order distribution of Y^ℓ, P_{Y^ℓ}, under certain conditions. Information measures, like entropies, conditional entropies, divergences, and mutual informations, will be denoted according to the conventional rules of the information theory literature, where it should be kept in mind that these measures may involve both ordinary and auxiliary random variables. For example, $I(\hat{U}^\ell; Y^\ell)$ and $H(Y^\ell|\hat{U}^\ell)$ are, respectively, the mutual information and the conditional entropy induced by the empirical distribution of $\hat{U}^\ell$, $P_{\hat{U}^\ell}$, and the conditional distribution, $P_{Y^\ell|\hat{U}^\ell}$, of the ordinary random vector, Y^ℓ, given the auxiliary random vector, $\hat{U}^\ell$. The conditional divergence, $D(Q_{Y|\hat{X}}\|P_{Y|\hat{X}}|P_{\hat{X}})$, will be understood to be given by

$$D(Q_{Y|\hat{X}}\|P_{Y|\hat{X}}|P_{\hat{X}}) = \sum_{x\in\mathcal{X}} P_{\hat{X}}(x) \sum_{y\in\mathcal{Y}} Q_{Y|\hat{X}}(y|x) \log \frac{Q_{Y|\hat{X}}(y|x)}{P_{Y|\hat{X}}(y|x)}, \tag{1}$$

where logarithms, here and throughout the sequel, will be understood to be taken to the base 2, unless specified otherwise.

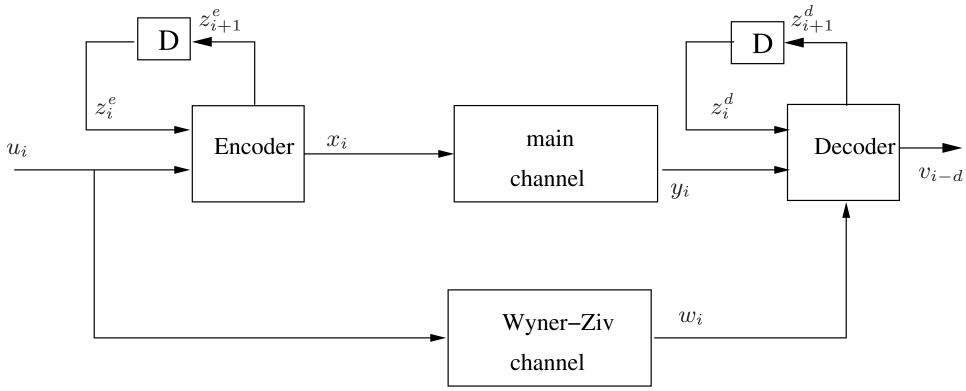

Figure 1: Source–channel with side information at the decoder according to the formal model description in Subsection 2.2. Both the encoder and the decoder are finite–state machines with an overall delay, d. The label "D", in each of the feedback loops of the encoder and the decoder, designates a device that introduces a delay of one time unit, thus passing from the next state (at time $i+1$) to the current state (at time i).

2.2 Problem Formulation

Referring to Fig. 1, let $\boldsymbol{u} = (u_1, u_2, \ldots)$ be a deterministic source sequence of symbols in a finite alphabet $\mathcal{U}$ of cardinality $|\mathcal{U}| = \alpha$. The sequence $\boldsymbol{u}$ is encoded using a periodically time–varying finite–state encoder, whose output is $\boldsymbol{x} = (x_1, x_2, \ldots)$, where $x_i \in \mathcal{X}$, $\mathcal{X}$ being another finite alphabet, of size $|\mathcal{X}| = \beta$. More precisely, the encoder obeys the following equations

$$t = i \bmod \ell \tag{2}$$

$$x_i = f_t(u_i, z_i^{\mathrm{e}}), \tag{3}$$

$$z_{i+1}^{\mathrm{e}} = g_t(u_i, z_i^{\mathrm{e}}), \tag{4}$$

where $i = 1, 2, \ldots$, z_i^{e} is the encoder state at time i, which takes values in a finite set, $\mathcal{Z}^{\mathrm{e}}$, of size s_{e}. The functions f_t and g_t are, respectively, the periodically time–varying *output function* and the *next–state function* of the encoder. The length of the period is ℓ. The sequence $\boldsymbol{x}$ is fed into a DMC, characterized by the single-letter transition probabilities $\{P_{Y|X}(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$, where $\mathcal{Y}$ is a finite alphabet of size $|\mathcal{Y}| = \gamma$. The channel output $\boldsymbol{y} = (y_1, y_2, \ldots)$ is fed into a periodically time–varying finite–state decoder, which is defined by

$$t = i \bmod \ell, \qquad i = 1, 2, \ldots \tag{5}$$

$$v_{i-d} = f'_t(w_i, y_i, z_i^{\mathrm{d}}), \qquad i = d+1, d+2, \ldots \tag{6}$$

$$z_{i+1}^{\rm d} \quad = \quad g_t'(w_i, y_i, z_i^{\rm d}), \qquad\qquad i = 1, 2, \ldots \tag{7}$$

where $z_i^{\rm d} \in \mathcal{Z}^{\rm d}$ is the decoder state at time i, $\mathcal{Z}^{\rm d}$ being a finite set of states of size $s_{\rm d}$, $w_i \in \mathcal{W}$ is the side information at time i, and v_{i-d} is the reconstructed version of the source sequence, delayed by d time units (d – positive integer). The reconstruction alphabet is $\mathcal{V}$ of size δ. The side information sequence, $\boldsymbol{w} = (w_1, w_2, \ldots)$, is generated from $\boldsymbol{u}$ by means of a DMC, characterized by a matrix of transition probabilities, $P_{W|U} = \{P_{W|U}(w|u),\ u \in \mathcal{U},\ w \in \mathcal{W}\}$. The functions f_t' and g_t' are, respectively, the output function and the next–state function of the decoder.

Let $V^n = (V_1, \ldots, V_n)$, $W^n = (W_1, \ldots, W_n)$, and $Y^n = (Y_1, \ldots, Y_n)$, designate random vectors pertaining to the variables $v^n = (v_1, \ldots, v_n)$, $w^n = (w_1, \ldots, w_n)$, and $y^n = (y_1, \ldots, y_n)$, respectively, where the randomness stems from the main channel and the W–Z channel. The vector u^n clearly does not have a stochastic counterpart. The same comment applies also to x^n since the encoder is assumed deterministic.

The objectives of the paper are the following: given the source sequence u^n, the channel, $P_{Y|X}$, and the W–Z channel $P_{W|U}$, the numbers of encoder and decoder states, $s_{\rm e}$ and $s_{\rm d}$, the period, ℓ, and the allowed delay, d, and given a single–letter distortion function $\rho : \mathcal{U} \times \mathcal{V} \to \mathbb{R}^+$, we wish to find non–trivial lower bounds to:

1. The expected distortion, $\frac{1}{n}\sum_{i=1}^n \boldsymbol{E}\{\rho(u_i, V_i)\}$, and
2. The probability of excess distortion, $\Pr\{\sum_{i=1}^n \rho(u_i, V_i) \geq nD\}$, where $D > 0$ is a constant larger than the best achievable normalized expected distortion.

In some instances of the problem, we will also discuss asymptotic achievability.

3 Background and Preliminary Results

Before moving forward to our main results, we need a few more definitions, as well as some background on the relevant results from [16]. We conclude this section with a preliminary result on the excess distortion probability in the case of an ordinary DMC and without source–related side information at the decoder.

Let ℓ divide n and consider the segmentation of all relevant sequences into n/ℓ non–overlapping blocks of length ℓ, that is,

$$u^n = (\boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{n/\ell-1}), \qquad \boldsymbol{u}_i = (u_{i\ell+1}, u_{i\ell+2}, \ldots, u_{i\ell+\ell}), \quad i = 0, 1, \ldots, n/\ell - 1, \tag{8}$$

and similar definitions for v^n, w^n, x^n, and y^n, where $v_{n-d+1}, v_{n-d+2}, \ldots, v_n$ (which are not yet reconstructed at time $t = n$) are defined as arbitrary symbols in $\mathcal{V}$. Let us define the empirical joint probability mass function

$$\begin{aligned} & P_{\hat{U}^\ell \hat{V}^\ell \hat{W}^\ell \hat{X}^\ell \hat{Y}^\ell \hat{Z}^{\mathrm{e}} \hat{Z}^{\mathrm{d}}}(u^\ell, v^\ell, w^\ell, x^\ell, y^\ell, z^{\mathrm{e}}, z^{\mathrm{d}}) \\ = \;& \frac{\ell}{n} \sum_{i=0}^{n/\ell-1} 1\{\boldsymbol{u}_i = u^\ell, \boldsymbol{v}_i = v^\ell, \boldsymbol{w}_i = w^\ell, \boldsymbol{x}_i = x^\ell, \boldsymbol{y}_i = y^\ell, z^{\mathrm{e}}_{i\ell+1} = z^{\mathrm{e}}, z^{\mathrm{d}}_{i\ell+1} = z^{\mathrm{d}}\}, \end{aligned} \tag{9}$$

where $1\{\ldots\}$ is the indicator function of the combination of events indicated in its argument. Clearly, since $P_{\hat{U}^\ell \hat{V}^\ell \hat{W}^\ell \hat{X}^\ell \hat{Y}^\ell \hat{Z}^{\mathrm{e}} \hat{Z}^{\mathrm{d}}}$ is a legitimate probability distribution, all the rules of manipulating information measures (the chain rule, conditioning reduces entropy, etc.) hold as usual. Marginal and conditional marginal distributions associated with subsets of the set of random variables, $(\hat{U}^\ell, \hat{V}^\ell, \hat{W}^\ell, \hat{X}^\ell, \hat{Y}^\ell, \hat{Z}^{\mathrm{e}}, \hat{Z}^{\mathrm{d}})$, which are derived from $P_{\hat{U}^\ell \hat{V}^\ell \hat{W}^\ell \hat{X}^\ell \hat{Y}^\ell \hat{Z}^{\mathrm{e}} \hat{Z}^{\mathrm{d}}}$, will be denoted using the conventional notation, for example, $P_{\hat{U}^\ell \hat{Y}^\ell}$ is the joint empirical distribution of $(\hat{U}^\ell, \hat{Y}^\ell)$, $P_{\hat{Y}^\ell|\hat{X}^\ell \hat{Z}^{\mathrm{e}}}$ is the conditional empirical distribution of $\hat{Y}^\ell$ given $(\hat{X}^\ell, \hat{Z}^{\mathrm{e}})$, and so on.

We now define the W-Z rate–distortion function [24] of the source $P_{\hat{U}^\ell}$ with respect to the *real* side–information channel $P_{W^\ell|\hat{U}^\ell}$ (as opposed to the empirical side–information channel, $P_{\hat{W}^\ell|\hat{U}^\ell}$) according to

$$R^{\mathrm{WZ}}_{\hat{U}^\ell|W^\ell}(D) = \frac{1}{\ell} \min\{I(\hat{U}^\ell; A) - I(W^\ell; A)\} \equiv \min \frac{1}{\ell} I(\hat{U}^\ell; A|W^\ell), \tag{10}$$

where both minima are taken over all conditional distributions, $\{P_{A|\hat{U}^\ell}\}$, such that $A \to \hat{U}^\ell \to W^\ell$ is a Markov chain and $\min_{\{G:\mathcal{A}\times\mathcal{W}^\ell\to\mathcal{V}^\ell\}} \boldsymbol{E}\{\rho(\hat{U}^\ell, G(A, W^\ell))\} \le \ell \cdot D$, where $\rho(u^\ell, v^\ell)$ is defined additively as $\sum_{i=1}^\ell \rho(u_i, v_i)$ and A is an auxiliary RV whose alphabet size is $|\mathcal{A}| = \alpha^\ell + 1$. It follows from these definitions that if $W^\ell \to \hat{U}^\ell \to A$ is a Markov chain, then

$$I(\hat{U}^\ell; A) - I(W^\ell; A) \equiv I(\hat{U}^\ell; A|W^\ell) \ge \ell \cdot R_{\hat{U}^\ell|W^\ell}[\Delta(\hat{U}^\ell|W^\ell, A)/\ell], \tag{11}$$

where

$$\Delta(\hat{U}^\ell|W^\ell, A) \stackrel{\Delta}{=} \min_{G:\mathcal{W}^\ell\times\mathcal{A}\to\mathcal{V}^\ell} \boldsymbol{E}\{\rho(\hat{U}^\ell, G(W^\ell, A))\}, \tag{12}$$

where $\mathcal{A}$ is the alphabet of A and the expectation is taken with respect to (w.r.t.) $P_{\hat{U}^\ell W^\ell A} = P_{\hat{U}^\ell A} \times P_{W^\ell|\hat{U}^\ell}$. Clearly, if W^ℓ is independent of $\hat{U}^\ell$, that is, $P_{W^\ell|\hat{U}^\ell}(\cdot|u^\ell)$ is the same for all $u^\ell \in \mathcal{U}^\ell$, then $R^{\mathrm{wz}}_{\hat{U}^\ell|W^\ell}(D)$ degenerates to the ordinary rate–distortion function, which will be denoted by $R_{\hat{U}^\ell}(D)$. In the sequel, we will also refer to the conditional rate–distortion, of $\hat{U}^\ell$ given W^ℓ, which will be denoted by $R_{\hat{U}^\ell|W^\ell}(D)$, where W^ℓ is available to both encoder and decoder. This function is also given by the minimum of $I(\hat{U}^\ell; A|W^\ell)/\ell$, except that the Markov condition is dropped [25]. The corresponding distortion–rate functions, $D^{\mathrm{wz}}_{\hat{U}^\ell|W^\ell}(R)$, $D_{\hat{U}^\ell}(R)$, and $D_{\hat{U}^\ell|W^\ell}(R)$, are the inverse functions of $R^{\mathrm{wz}}_{\hat{U}^\ell|W^\ell}(D)$, $R_{\hat{U}^\ell}(D)$, and $R_{\hat{U}^\ell|W^\ell}(D)$, respectively.

For the given channel, $P_{Y|X}$, we denote by $C_{P_{Y|X}}(\Gamma)$ the channel capacity with a transmission cost constraint, $\sum_{i=1}^n \boldsymbol{E}\{\phi(X_i)\} \le n\Gamma$ ($\phi(\cdot)$ being the single–letter transmission cost function), that is

$$C_{P_{Y|X}}(\Gamma) = \max_{P_X:\ \boldsymbol{E}\{\phi(X)\} \le \Gamma} I(X;Y). \tag{13}$$

When the channel, $P_{Y|X}$, is clear from the context, the subscript "$P_{Y|X}$" will be omitted, and the notation will be simplified to $C(\Gamma)$.

In [16], we considered the simpler case without decoder side information, $\boldsymbol{w}$, related to the source, and where only the decoder is limited to s states. One of the main results of [16] (in particular, Theorem 1 therein) is a lower bound to the expected distortion, which has the following form:

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{E}\{\rho(u_i, V_i)\} \ge D_{\hat{U}^\ell}\left(C(\Gamma) + \zeta(s,d,\ell) + \epsilon(\ell,n)\right), \tag{14}$$

where

$$\epsilon(\ell,n) \stackrel{\triangle}{=} \frac{(\alpha\beta)^\ell \log\gamma}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right), \tag{15}$$

and $\zeta(s,d,\ell)$ is a certain function[1] with the property $\lim_{\ell\to\infty}\zeta(s,d,\ell) = 0$. As discussed in [16], it is interesting that the distortion bound depends on u^n only via its ℓ–th order empirical distribution, $P_{\hat{U}^\ell}$, where, as defined above, ℓ is length of the period. It is also discussed in that work that the term $\zeta(s,d,\ell)$, on the right–hand side, can be thought of as an "extra capacity" term, that is induced by the memory of the encoding–decoding system (encapsulated in the state) and the allowed delay, but its effect is diminished when ℓ is chosen large. The bound can then be asymptotically approached

[1] The exact form of this function is immaterial for the purpose of this discussion. In fact, the formula of $\zeta(s,d,\ell)$, given in [16] is somewhat imprecise, and so, we both correct and extend it in this work. Nevertheless, the property $\lim_{\ell\to\infty}\zeta(s,d,\ell) = 0$ remains valid.

by separate source– and channel coding, using long block codes (see also the achievability scheme described in detail in the discussion of Section 4 below). On the other hand, by letting ℓ grow, one also affects the distortion–rate function, $D_{\hat{U}^\ell}(\cdot)$, and so, the overall effect of ℓ is not trivial to assess in general.

We now state our preliminary result on the excess distortion probability for the case of DMC and without any side information.

Theorem 1 *Assume that $\rho_{\max} = \max_{u,v} \rho(u,v) < \infty$. Then, under the assumptions of [16], for a given u^n, an arbitrary encoder and a finite–state decoder with s states and overall delay d,*

$$\begin{aligned} Pr\left\{\sum_{i=1}^n \rho(u_i, V_i) \geq nD\right\} \quad &\geq \quad \sup_{\Delta>0}\left[\frac{\Delta}{\rho_{\max} - D} - o(n)\right] \times \\ &\quad\; \exp\left\{-(n+d)E_{sp}\left[R_{\hat{U}^\ell}(D+\Delta) - \zeta(s,d,\ell) - \epsilon(\ell,n)\right]\right\}, \qquad (16) \end{aligned}$$

where $E_{sp}(R)$ is the sphere–packing exponent of the channel, i.e.,

$$E_{sp}(R) = \sup_{Q_X} \inf_{\{Q_{Y|X}:\; I_Q(X;Y) \leq R\}} D(Q_{Y|X} \| P_{Y|X} | Q_X), \tag{17}$$

with $I_Q(X;Y)$ denoting the mutual information induced by $Q_X \times Q_{Y|X}$.

As a simple conclusion from this theorem, we have that

$$\liminf_{n\to\infty} \frac{1}{n} \log\left[\Pr\left\{\sum_{i=1}^n \rho(u_i, V_i) \geq nD\right\}\right] \geq -E_{\rm sp}\left[R_{\hat{U}^\ell}(D+0) - \zeta(s,d,\ell)\right], \tag{18}$$

where

$$R_{\hat{U}^\ell}(D+0) \stackrel{\Delta}{=} \lim_{\Delta\downarrow 0} R_{\hat{U}^\ell}(D+\Delta). \tag{19}$$

Discussion.

First, observe that the "extra capacity" term, $\zeta(s,d,\ell)$, plays a role here too. This time, it appears in the form of an effective rate reduction in the argument of the sphere–packing error exponent. But once again, if ℓ is very large while s and d are fixed, this term becomes insignificant. In this case, as long as D is smaller than $D_{\hat{U}^\ell}(R_{\rm crit})$, $R_{\rm crit}$ being the critical rate [7] of the channel (and assuming $D_{\hat{U}^\ell}(R_{\rm crit}) > 0$), the bound is asymptotically achievable using long blocks (of size $n \gg \ell$), by rate–distortion coding (based on the type covering lemma) in the superalphabet of ℓ–vectors, followed by channel coding, as in Csiszár's works, [4] and [5, Theorem 2]. As in those references, strictly speaking, this is not quite considered separate source– and channel coding, because there is a certain linkage between the channel code design and the source: Each type class of source sequences (in the level of ℓ-blocks) is mapped into a channel sub–code at rate $R_{\hat{U}^\ell}(D)$ (approximately), and the corresponding channel codewords are of the type, $P_{\hat{X}}$, that achieves the maximum sphere–packing exponent at that particular rate.

As a side remark, speaking of Csiszár's source–channel error exponents, [4] and [5], it is interesting to relate Theorem 1 above to its purely probabilistic counterpart. The proof of Theorem 1 above is based on a change–of–measure argument. Since only the channel is probabilistic in our setting, the upper bound on the exponent includes only a channel–related term, which is the channel's sphere–packing exponent. Applying a similar line of thought in the purely probabilistic case, we have to change measures for both the source and the channel, and so, we end up minimizing the sum of two divergence terms, i.e.,

$$\min_{\{(Q_U,Q_{Y|X}):\ R_{Q_U}(D)\geq I_Q(X;Y)\}}\{D(Q_U\|P_U)+D(Q_{Y|X}\|P_{Y|X}|Q_X)\}, \tag{20}$$

where P_U is the memoryless source and $P_{Y|X}$ is the memoryless channel. This upper bound on the joint source–channel exponent can be further upper bounded by arbitrarily selecting a positive real R and arguing that

$$\begin{aligned}
& \min_{\{(Q_U,Q_{Y|X}):\ R_{Q_U}(D)\geq I_Q(X;Y)\}}\{D(Q_U\|P_U)+D(Q_{Y|X}\|P_{Y|X}|Q_X)\} \\
\leq\ & \min_{\{(Q_U,Q_{Y|X}):\ R_{Q_U}(D)\geq R\geq I_Q(X;Y)\}}\{D(Q_U\|P_U)+D(Q_{Y|X}\|P_{Y|X}|Q_X)\} \\
=\ & \min_{\{(Q_U:\ R_{Q_U}(D)\geq R\}} D(Q_U\|P_U)+\min_{\{Q_{Y|X}:\ I_Q(X;Y)\leq R\}} D(Q_{Y|X}\|P_{Y|X}|Q_X) \\
\leq\ & F(R,D,U)+E_{\rm sp}(R),
\end{aligned} \tag{21}$$

where $F(R,D,U)$ is Marton's source coding exponent [14] for the memoryless source $U\sim P_U$. Since this argument is applicable to any value of R, the tightest upper bound is obtained by minimizing over R, namely, the resulting upper bound on the exponent is

$$\min_R[F(R,D,U)+E_{\rm sp}(R)], \tag{22}$$

which coincides with Csiszár's upper bound [5, Theorem 4] to the best achievable excess–distortion exponent. This argument, however, is quite different from the one used in [5], which in turn is based on the list–decoding argument of Shannon, Gallager and Berlekamp [22], that originally, sets the stage for the straight–line bound [7, Theorem 5.8.2].

Returning to the main topic of this work, the remaining part of this section is devoted to the proof of Theorem 1.

Proof of Theorem 1. Let $\Delta > 0$ be arbitrarily small and let $Q_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^n Q_{Y|X}(y_i|x_i)$ be an auxiliary DMC such that

$$\ell \cdot R_{\hat{U}^\ell}(D+\Delta) \geq \ell \cdot I_Q(\hat{X};Y) + \ell\zeta(s,d,\ell) + \ell\epsilon(\ell,n) = \ell[I_Q(\hat{X};Y)+\lambda], \tag{23}$$

where $\lambda \stackrel{\Delta}{=} \zeta(s,d,\ell)+\epsilon(\ell,n)$, and where the empirical channel input distribution $P_{\tilde{X}}$ is induced by the encoder output, x^{n+d}. Since $Q_{Y|X}$ is assumed memoryless, we have the following relationship between $I_Q(\tilde{X};Y)$ and $I_Q(\tilde{X}^\ell;Y^\ell)$ ($\tilde{X}^\ell$ being the random vector induced by the empirical distribution of ℓ–blocks, extracted from x^{n+d}, assuming that ℓ divides $n+d$):

$$\begin{aligned}
\frac{I_Q(\tilde{X}^\ell;Y^\ell)}{\ell} &\leq \frac{1}{\ell}\sum_{j=1}^{\ell} I_Q(\tilde{X}_j;Y_j) && (24)\\
&= I(\tilde{X}_J;Y_J|J) && (25)\\
&= H(Y_J|J) - H(Y_J|\tilde{X}_J,J) && (26)\\
&\leq H(Y_J) - H(Y_J|\tilde{X}_J,J) && (27)\\
&= H(Y_J) - H(Y_J|\tilde{X}_J) && (28)\\
&= I_Q(\tilde{X}_J;Y_J) && (29)\\
&= I_Q(\tilde{X};Y), && (30)
\end{aligned}$$

where $\tilde{X}_j$ is the random variable derived from the the j–th marginal of $P_{\tilde{X}^\ell}$, $j=1,2,\ldots,\ell$, J is an integer random variable, uniformly distributed over $\{1,2,\ldots,\ell\}$, and where we have used the Markovity of the chain $J \to \hat{X}_J \to Y_J$, and the identities $\tilde{X}_J = \tilde{X}$, $Y_J = Y$, which follow from the fact that $P_{\tilde{X}} = \frac{1}{\ell}\sum_{j=1}^{\ell} P_{\tilde{X}_j} = P_{\tilde{X}_J}$. Therefore, if Q satisfies (23), it must also satisfy

$$\ell \cdot R_{\hat{U}^\ell}(D+\Delta) \geq I_Q(\tilde{X}^\ell;Y^\ell) + \ell\lambda. \tag{31}$$

According to the above cited Theorem 1 of [16], for such a channel, the expected distortion under Q, denoted D_Q, must be lower bounded by

$$D_Q \stackrel{\Delta}{=} \boldsymbol{E}_Q\left\{\frac{1}{n}\sum_{i=1}^{n}\rho(u_i,V_i)\right\} \geq D_{\hat{U}^\ell}\left(\frac{I_Q(\tilde{X}^\ell;Y^\ell)}{\ell}+\lambda\right) \geq D+\Delta. \tag{32}$$

Denoting the event $\mathcal{E} = \{y^{n+d} : \ \rho(u^n, v^n)) > nD\}$, we have

$$
\begin{aligned}
D_Q &= \frac{1}{n}\boldsymbol{E}_Q\{\rho(u^n, V^n))\} \\
&= \frac{1}{n}\boldsymbol{E}_Q\{\rho(u^n, V^n)\cdot 1\{\mathcal{E}\}\} + \frac{1}{n}\boldsymbol{E}_Q\{\rho(u^n, V^n))\cdot 1\{\mathcal{E}^{\mathrm{c}}\}\} \\
&\le \frac{1}{n}\cdot Q(\mathcal{E})\cdot n\rho_{\max} + \frac{1}{n}[1 - Q(\mathcal{E})]\cdot nD \\
&= [1 - Q(\mathcal{E})]\cdot D + Q(\mathcal{E})\cdot \rho_{\max}, \qquad (33)
\end{aligned}
$$

which implies that

$$
Q(\mathcal{E}) \ge \frac{D_Q - D}{\rho_{\max} - D} \ge \frac{D + \Delta - D}{\rho_{\max} - D} = \frac{\Delta}{\rho_{\max} - D}. \qquad (34)
$$

Now, for a given, arbitrarily small $\epsilon_0 > 0$, let us define

$$
\mathcal{T} = \left\{ y^{n+d} : \ \sum_{i=1}^{n+d} \ln\frac{Q_{Y|X}(y_i|x_i)}{P_{Y|X}(y_i|x_i)} \le (n+d)[D(Q_{Y|X}\|P_{Y|X}|P_{\tilde{X}}) + \epsilon_0] \right\}. \qquad (35)
$$

Now,

$$
\begin{aligned}
\Pr\{\rho(u^n, V^n) \ge nD\} &= \sum_{y^{n+d}\in\mathcal{E}} P_{Y^{n+d}|X^{n+d}}(y^{n+d}|x^{n+d}) \\
&\ge \sum_{y^{n+d}\in\mathcal{E}\cap\mathcal{T}} P_{Y^{n+d}|X^{n+d}}(y^{n+d}|x^{n+d}) \\
&= \sum_{y^{n+d}\in\mathcal{E}\cap\mathcal{T}} Q_{Y^{n+d}|X^{n+d}}(y^{n+d}|x^{n+d})\cdot\exp\left\{-\sum_{i=1}^{n+d}\log\frac{Q_{Y|X}(y_i|x_i)}{P_{Y|X}(y_i|x_i)}\right\} \\
&\ge \sum_{y^{n+d}\in\mathcal{E}\cap\mathcal{T}} Q_{Y^{n+d}|X^{n+d}}(y^{n+d}|x^{n+d})\cdot\exp\left\{-(n+d)[D(Q_{Y|X}\|P_{Y|X}|P_{\tilde{X}}) + \epsilon_0]\right\} \\
&= \exp\left\{-(n+d)[D(Q_{Y|X}\|P_{Y|X}|P_{\tilde{X}}) + \epsilon_0]\right\}\cdot Q(\mathcal{E}\cap\mathcal{T}) \\
&\ge \exp\left\{-(n+d)[D(Q_{Y|X}\|P_{Y|X}|P_{\tilde{X}}) + \epsilon_0]\right\}\cdot [Q(\mathcal{E}) - Q(\mathcal{T}^{\mathrm{c}})] \\
&\ge \exp\left\{-(n+d)[D(Q_{Y|X}\|P_{Y|X}|P_{\tilde{X}}) + \epsilon_0]\right\}\cdot \left[\frac{\Delta}{\rho_{\max} - D} - o(n)\right], \qquad (36)
\end{aligned}
$$

where we have used the fact that $Q(\mathcal{T}^{\mathrm{c}}) = o(n)$ for every $\epsilon_0 > 0$, by the weak law of large numbers. Since $Q_{Y|X}$ is an arbitrary channel that satisfies (23), we have

$$
\begin{aligned}
\Pr\{\rho(u^n, V^n) \ge nD\} &\ge \left[\frac{\Delta}{\rho_{\max} - D} - o(n)\right]\times \\
&\quad \exp\left\{-(n+d)\inf_{\{Q_{Y|X}:\ I_Q(\hat{X};Y)\le R_{\hat{U}^\ell}(D+\Delta)-\lambda\}} D(Q_{Y|X}\|P_{Y|X}|P_{\tilde{X}}) + \epsilon_0\right\} \\
&\ge \left[\frac{\Delta}{\rho_{\max} - D} - o(n)\right]\cdot\exp\left\{-(n+d)\left(E_{\mathrm{sp}}[R_{\hat{U}^\ell}(D+\Delta) - \lambda] + \epsilon_0\right)\right\}, \qquad (37)
\end{aligned}
$$

which completes the proof of Theorem 1 by the arbitrariness of $\epsilon_0 > 0$.

4 Side Information at the Decoder

Consider again the setting described in Section 2 and depicted in Fig. 1. Our first result is the following:

Theorem 2 *Assume that $\rho_{\max} = \max_{u,v} \rho(u, v) < \infty$. For a given u^n, any finite–state encoder with s_e states, and any finite–state decoder with s_d states, both having a period of length ℓ and an overall delay d,*

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{E}\{\rho(u_i, V_i)\} \geq D^{WZ}_{\hat{U}^\ell|W^\ell}\left(C(\Gamma) + \frac{\log s_d}{\ell} + \frac{s_e \alpha^\ell \log \gamma}{\sqrt{n}} + o(n)\right) - \frac{\rho_{\max} d}{\ell}. \tag{38}$$

Discussion.

Similarly as in Theorem 1, the distortion lower bound depends on the source sequence, u^n, only via its empirical distribution from the order that corresponds to the period, ℓ. We also observe that the numbers of states, s_{e}, s_{d} and the allowed delay, d, take parts in the lower bound in different ways. The first two play the role of 'excess capacity', whereas the latter serves in a term of distortion reduction. But this difference is not really crucial, because excess rate and reduced distortion are two faces of the same coin. Indeed, one of technical issues in the proof of Theorem 2 below (which was not handled perfectly rigorously in the parallel derivation in [16]), evolves around the following question: how can one assess the effect of the decoder state contribution in estimating the source (and thereby reducing the distortion relative to the absence of the state), in order to bound the distortion in terms of W-Z block code performance. In other words, the question is: how to obtain a lower bound in terms of block codes, where no state carries information over from block to block? As will be seen in the proof below, the idea is that since the decoder state cannot carry more than $\log s_{\mathrm{d}}$ information bits about the source, its effect cannot be better than that of adding an excess rate of $\Delta R = (\log s_{\mathrm{d}})/\ell$ to the corresponding W-Z encoder. This demonstrates clearly the point that distortion reduction can be traded with excess rate.

Another important observation is that our bound depends very differently on s_{e} and s_{d}. The dependence upon s_{e}, although linear rather than logarithmic, is considerably weaker, as it vanishes as soon as the limit $n \to \infty$ is taken, whereas the dependence on s_{d} 'survives' the limit $n \to \infty$ and vanishes only after the limit $\ell \to \infty$ is taken. In other words, at least as far as the distortion lower bound in concerned, the number of states of the decoder is a much more important resource than the number of states of the encoder. We find this asymmetric behavior interesting and not trivial. A partial intuitive explanation to this 'discrimation' between encoder memory and decoder memory is the following: While at the encoder, the state can only help to drive the input to achieve capacity, at the encoder, on the other hand, the state may play an important role in helping to estimate the source by exploiting correlations with past source blocks, if exist. It should also be pointed out that in terms of the dependence on s_{d}, our new lower bound is tighter than those of both [16] and [27] (in spite of the fact that we consider here a more general setting of side information at the decoder): the coefficient in front of the extra-capacity term, $(\log s_{\mathrm{d}})/\ell$, is reduced from 2 (in both [16] and [27]) to 1 here.

For very large ℓ (compared to d and $\log s_{\mathrm{d}}$), the lower bound can be asymptotically approached by separate source- and channel coding: W-Z coding in the superalphabet of ℓ–vectors, followed by channel coding. Specifically, given a long block of source sequence, u^n, of length $n \gg \alpha^\ell$, compute the empirical distribution, $P_{\hat{U}^\ell}$, and construct a W–Z code for the 'source' $\{P_{\hat{U}^\ell}(u^\ell),\ u^\ell \in \mathcal{U}^\ell\}$ and the side information channel $P_{W^\ell|\hat{U}^\ell}$, for a distortion level tuned such that $R^{\mathrm{WZ}}_{\hat{U}^\ell|W^\ell}(D) \le C(\Gamma) - 2\epsilon$ for some prescribed, arbitrarily small $\epsilon > 0$. Append to the W-Z bitstream a header of length $\lceil \log(n/\ell + 1)^{\alpha^\ell - 1} \rceil$ in order to transmit to the decoder a description of the empirical distribution, $\{P_{\hat{U}^\ell}(u^\ell),\ u^\ell \in \mathcal{U}^\ell\}$. This is necessary for the receiver to apply the decoder corresponding to the encoder of that source. If n is large enough compared to $\alpha^\ell \log n$, then

$$\frac{\lceil \log(n/\ell + 1)^{\alpha^\ell - 1} \rceil}{n} \le \frac{(\alpha^\ell - 1)\log(n/\ell + 1) + 1}{n} \le \epsilon, \tag{39}$$

and the total channel coding rate is below $C(\Gamma) - \epsilon$. By applying a good channel code for this rate, the lower bound to the distortion is essentially achieved, similarly as in traditional separate source- and channel coding.

Block encoders of length n can be thought of finite–state devices with delay n and the same comment applies to block decoders. While the numbers of states of this block encoder and decoder are much larger than s_{e} and s_{d}, respectively, the gaps between the converse and the achievability bounds shrinks in the asymptotic limit where n tends to infinity and then ℓ tends to infinity. The above described achievability scheme is similar in spirit to those of [27] and [28]. Note that this separation–based scheme asymptotically meets the lower bound in spite of the fact the W-Z channel violates the Markov structure of traditional communication system, $\boldsymbol{u} \to \boldsymbol{x} \to \boldsymbol{y} \to \boldsymbol{v}$. This is coherent with the analogous behavior in the purely probabilistic setting [20], and moreover, even if the DMC is replaced by the Shannon channel. In the last part of the next section, we will refer to this channel.

Finally, on the basis of Theorem 2, and similarly as in Theorem 1, one can easily derive a lower bound on the excess distortion probability for the case of decoder side information, considered here. This time, however, the change of measures should involve, not only the main channel, $P_{Y|X}$, as before, but also the W-Z channel, $P_{W|\hat{U}}$. The resulting exponential lower bound would be of the form,

$$\begin{aligned}\Pr\left\{\sum_{i=1}^{n}\rho(u_i,V_i)\geq nD\right\} \quad \geq \quad & \sup_{\Delta>0}\left[\frac{\Delta}{\rho_{\max}-D}-o(n)\right]\times \\ & \exp\left\{-(n+d)\max_{Q_{\hat{X}}}\min\left[D(Q_{W|\hat{U}}\|P_{W|\hat{U}}|P_{\hat{U}})+\right.\right. \\ & \left.\left.D(Q_{Y|\hat{X}}\|P_{Y|\hat{X}}|Q_{\hat{X}})+2\epsilon_0\right]\right\},\end{aligned} \tag{40}$$

where the minimum is over all pairs $\{(Q_{W|\hat{U}},Q_{Y|\hat{X}})\}$ such that

$$R^{\mathrm{WZ}}_{\hat{U}^\ell,Q_{W^\ell|\hat{U}^\ell}}\left(D+\frac{\rho_{\max}d}{\ell}+\Delta\right)\geq I_Q(\hat{X};Y)+\frac{\log s_{\mathrm{d}}}{\ell}+\frac{s_{\mathrm{e}}\alpha^\ell\log\gamma}{\sqrt{n}}+o(n), \tag{41}$$

where $R^{\mathrm{WZ}}_{\hat{U}^\ell,Q_{W^\ell|\hat{U}^\ell}}(\cdot)$ is the W-Z rate–distortion function of $P_{\hat{U}^\ell}$ with the W-Z channel, $Q_{W^\ell|\hat{U}^\ell}$. Similarly as before, the bound is asymptotically achievable following the same line of thought as in [4] and [5], for $D < D^{\mathrm{WZ}}_{\hat{U}^\ell|W^\ell}(R_{\mathrm{crit}})$, where R_{crit} is the critical rate of the channel, provided that $D^{\mathrm{WZ}}_{\hat{U}^\ell|W^\ell}(R_{\mathrm{crit}}) > 0$.

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

Proof of Theorem 2. Owing to the encoder–decoder model (2), (5), it is clear that $x_{i\ell+1}^{i\ell+\ell}$ is a deterministic function of $z^{\mathrm{e}}_{i\ell+1}$ and $u_{i\ell+1}^{i\ell+\ell}$, as

$$\begin{aligned} x_{i\ell+1} &= f_1(u_{i\ell+1},z^{\mathrm{e}}_{i\ell+1}), \\ x_{i\ell+2} &= f_2(u_{i\ell+2},z^{\mathrm{e}}_{i\ell+2}) = f_2(u_{i\ell+2},g_1(u_{i\ell+1},z^{\mathrm{e}}_{i\ell+1})) \\ &\cdots \\ x_{i\ell+\ell} &= f_0(u_{i\ell+\ell},z^{\mathrm{e}}_{i\ell+\ell}). \end{aligned} \tag{42}$$

Accordingly, we denote

$$x_{i\ell+1}^{i\ell+\ell} = q(z_{i\ell+1}^{\mathrm{e}}, u_{i\ell+1}^{i\ell+\ell}), \tag{43}$$

where $q : \mathcal{Z}^{\mathrm{e}} \times \mathcal{U}^{\ell} \to \mathcal{X}^{\ell}$. Likewise, $v_{i\ell+1}^{i\ell+\ell-d}$ is a deterministic function of $z_{i\ell+1}^{\mathrm{d}}$, $y_{i\ell+1}^{i\ell+\ell}$ and $w_{i\ell+1}^{i\ell+\ell}$, as

$$\begin{aligned}
v_{i\ell+1} &= f'_{d+1}(w_{i\ell+d+1}, y_{i\ell+d+1}, z_{i\ell+d+1}^{\mathrm{d}}), \quad z_{i\ell+d+1}^{\mathrm{d}} \text{ being a function of } w_{i\ell+1}^{i\ell+d},\ y_{i\ell+1}^{i\ell+d} \text{ and } z_{i\ell+1}^{\mathrm{d}} \\
v_{i\ell+2} &= f'_{d+2}(w_{i\ell+d+2}, y_{i\ell+d+2}, z_{i\ell+d+2}^{\mathrm{d}}) \\
&\ldots \\
v_{i\ell+\ell-d} &= f'_{0}(w_{i\ell+\ell}, y_{i\ell+\ell}, z_{i\ell+\ell}^{\mathrm{d}}).
\end{aligned} \tag{44}$$

Accordingly, we denote

$$v_{i\ell+1}^{i\ell+\ell-d} = m(z_{i\ell+1}^{\mathrm{d}}, w_{i\ell+1}^{i\ell+\ell}, y_{i\ell+1}^{i\ell+\ell}), \tag{45}$$

where $m : \mathcal{Z}^{\mathrm{d}} \times \mathcal{W}^{\ell} \times \mathcal{Y}^{\ell} \to \mathcal{V}^{\ell-d}$. The proof of the theorem is based on deriving both a lower bound and an upper bound to the expected empirical conditional mutual information, $\boldsymbol{E}\{I(\hat{U}^{\ell}; \hat{Y}^{\ell})\}$, which applies to any finite–state encoder and any finite–state decoder with an overall delay d.

As for an upper bound, we have following.

$$\begin{aligned}
\boldsymbol{E}\{I(\hat{U}^{\ell}; \hat{Y}^{\ell})\} &\le \boldsymbol{E}\{I(\hat{U}^{\ell}, \hat{Z}^{\mathrm{e}}; \hat{Y}^{\ell})\} \\
&= \boldsymbol{E}\{H(\hat{Y}^{\ell})\} - \boldsymbol{E}\{H(\hat{Y}^{\ell}|\hat{U}^{\ell}, \hat{Z}^{\mathrm{e}})\} \\
&\le H(Y^{\ell}) - \boldsymbol{E}\{H(\hat{Y}^{\ell}|\hat{U}^{\ell}, \hat{Z}^{\mathrm{e}})\},
\end{aligned} \tag{46}$$

where the last inequality follows from the concavity of the entropy function. In Appendix A, we prove the following inequality:

$$\boldsymbol{E}\{H(\hat{Y}^{\ell}|\hat{U}^{\ell}, \hat{Z}^{\mathrm{e}})\} \ge H(Y^{\ell}|\hat{X}^{\ell}) - \ell \cdot \Delta_1(s_{\mathrm{e}}, \ell, n), \tag{47}$$

where

$$\Delta_1(s_{\mathrm{e}}, \ell, n) \stackrel{\Delta}{=} \frac{s_{\mathrm{e}} \alpha^{\ell} \log \gamma}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{48}$$

Combining this with eq. (46), we get

$$\begin{aligned}
\boldsymbol{E}\{I(\hat{U}^{\ell}; \hat{Y}^{\ell})\} &\le H(Y^{\ell}) - H(Y^{\ell}|\hat{X}^{\ell}) + \ell \cdot \Delta_1(s_{\mathrm{e}}, \ell, n) \\
&= I(\hat{X}^{\ell}; Y^{\ell}) + \ell \cdot \Delta_1(s_{\mathrm{e}}, \ell, n) \\
&\le \ell \cdot [C(\Gamma) + \Delta_1(s_{\mathrm{e}}, \ell, n)].
\end{aligned} \tag{49}$$

Note in passing that in the last step of (46) one could use a tighter upper bound: instead of maximizing over all $\{P_{X^\ell}\}$ that comply with the transmission cost constraint, we could have also maximized over all $\{P_{X^\ell}\}$ that maintain the same empirical single–letter marginal, $P_{\hat{X}}$ (refer to eq. (24)). While this fact is immaterial for the expected distortion lower bound, it will be important when it comes to the lower bound on the excess–distortion probability.

To derive a lower bound to $\boldsymbol{E}\{I(\hat{U}^\ell;\hat{Y}^\ell)\}$, we first underestimate $I(\hat{U}^\ell;\hat{Y}^\ell)$ without taking the expectation.

$$\begin{aligned}
I(\hat{U}^\ell,\hat{Y}^\ell) &= I(\hat{U}^\ell,W^\ell;\hat{Y}^\ell) \\
&\geq I(\hat{U}^\ell,W^\ell;\hat{Y}^\ell) - I(W^\ell;\hat{Y}^\ell) \\
&= I(\hat{U}^\ell;\hat{Y}^\ell|W^\ell) \\
&\geq \ell\cdot R^{\mathrm{wz}}_{\hat{U}^\ell|W^\ell}[\Delta(\hat{U}^\ell|\hat{Y}^\ell,W^\ell)/\ell],
\end{aligned} \tag{50}$$

where the underlying joint distribution of $(\hat{U}^\ell,\hat{Y}^\ell,W^\ell)$ is assumed $P_{\hat{U}^\ell,\hat{Y}^\ell}\times P_{W^\ell|\hat{U}^\ell}$. The first equality follows from the fact that $W^\ell\to\hat{U}^\ell\to\hat{Y}^\ell$ forms a Markov chain. Consider now the joint distribution $P_{\hat{U}^\ell W^\ell\hat{Y}^\ell\hat{Z}^{\mathrm{d}}} = P_{\hat{U}^\ell\hat{Y}^\ell\hat{Z}^{\mathrm{d}}}\times P_{W^\ell|\hat{U}^\ell}$, where $\hat{Z}^{\mathrm{d}}$ designates the decoder state whose alphabet size is s_{d}. We argue that no matter what this distribution may be, the best resulting distortion in estimating $\hat{U}^\ell$ from $(W^\ell,\hat{Y}^\ell,\hat{Z}^{\mathrm{d}})$, that is, $\Delta(\hat{U}^\ell|W^\ell,\hat{Y}^\ell,\hat{Z}^{\mathrm{d}})$, cannot be better than the minimum achievable distortion when $\hat{Z}^{\mathrm{d}}$ is replaced by another random variable, Z^{d}_*, of the same alphabet size s_{d}, that is given by a deterministic function of $\hat{U}^\ell$ (see also [11, Proof of the converse to Theorem 6] for the use of a similar idea, albeit in a very different context). Indeed,

$$\begin{aligned}
\Delta(\hat{U}^\ell|W^\ell,\hat{Y}^\ell,\hat{Z}^{\mathrm{d}}) &= \min_G \sum_{u^\ell} P_{\hat{U}^\ell}(u^\ell)\sum_{z^{\mathrm{d}}} P_{\hat{Z}^{\mathrm{d}}|\hat{U}^\ell}(z^{\mathrm{d}}|u^\ell)\times \\
&\quad \sum_{w^\ell,y^\ell} P_{W^\ell\hat{Y}^\ell|\hat{U}^\ell\hat{Z}^{\mathrm{d}}}(w^\ell,y^\ell|u^\ell,z^{\mathrm{d}})\rho(u^\ell,G(w^\ell,y^\ell,z^{\mathrm{d}}))
\end{aligned} \tag{51}$$

is minimized by the conditional distribution, $P_{\hat{Z}^{\mathrm{d}}|\hat{U}^\ell}$, that puts all its mass on

$$z^{\mathrm{d}}_*(u^\ell) = \arg\min_{z^{\mathrm{d}}} c(u^\ell,z^{\mathrm{d}}), \tag{52}$$

where

$$c(u^\ell,z^{\mathrm{d}}) \stackrel{\Delta}{=} \sum_{w^\ell,y^\ell} P_{W^\ell\hat{Y}^\ell|\hat{U}^\ell\hat{Z}^{\mathrm{d}}}(w^\ell,y^\ell|u^\ell,z^{\mathrm{d}})\rho(u^\ell,G(w^\ell,y^\ell,z^{\mathrm{d}})). \tag{53}$$

Since $Z_*^{\rm d}$ is a deterministic function of $\hat{U}^\ell$, it is available to the encoder. Consider now a coding scheme that transmits $I(\hat{U}^\ell;\hat{Y}^\ell|W^\ell)$ bits per ℓ–vector using a Wyner–Ziv code (with W^ℓ serving as side information at the decoder), plus additional $\log s_{\rm d}$ bits to transmit $Z_*^{\rm d}$ as additional information on $\hat{U}^\ell$. The overall code–length of $I(\hat{U}^\ell;\hat{Y}^\ell|W^\ell)+\log s_{\rm d}$ bits cannot be smaller than that of the best Wyner–Ziv code that makes $(W^\ell,\hat{Y}^\ell,Z_*^{\rm d})$ available to the decoder, with an extra rate of $\log s_{\rm d}$ bits, as the latter can potentially exploit the statistical dependence between $\hat{U}^\ell$ and $Z_*^{\rm d}$. In other words, the point $((I(\hat{U}^\ell;\hat{Y}^\ell|W^\ell)+\log s_{\rm d})/\ell, \Delta(\hat{U}^\ell|W^\ell,\hat{Y}^\ell,Z_*^{\rm d})/\ell)$ is an achievable point on the rate–distortion plane, which implies that

$$I(\hat{U}^\ell;\hat{Y}^\ell|W^\ell)+\log s_{\rm d} \ge \ell\cdot R^{\rm wz}_{\hat{U}^\ell|W^\ell}[\Delta(\hat{U}^\ell|W^\ell,\hat{Y}^\ell,Z_*^{\rm d})/\ell] \ge \ell\cdot R^{\rm wz}_{\hat{U}^\ell|W^\ell}[\Delta(\hat{U}^\ell|W^\ell,\hat{Y}^\ell,\hat{Z}^{\rm d})/\ell], \quad (54)$$

where the last equality follows from $\Delta(\hat{U}^\ell|W^\ell,\hat{Y}^\ell,Z_*^{\rm d}) \le \Delta(\hat{U}^\ell|W^\ell,\hat{Y}^\ell,\hat{Z}^{\rm d})$ and the non–increasing monotonicity of the Wyner–Ziv rate–distortion function.

Let us now focus on the quantity $\Delta(\hat{U}^\ell|W^\ell,\hat{Y}^\ell,\hat{Z}^{\rm d})/\ell$:

$$\begin{aligned}
\frac{1}{\ell}\Delta(\hat{U}^\ell|W^\ell,\hat{Y}^\ell,\hat{Z}^{\rm d}) &= \min_G \frac{1}{\ell}\sum_{u^\ell,w^\ell,y^\ell,z^{\rm d}} P_{\hat{U}^\ell W^\ell \hat{Y}^\ell \hat{Z}^{\rm d}}(u^\ell,w^\ell,y^\ell,z^{\rm d})\rho(u^\ell,G(w^\ell,y^\ell,z^{\rm d}))\\
&\le \frac{1}{\ell}\sum_{u^\ell,w^\ell,y^\ell,z^{\rm d}} P_{\hat{U}^\ell W^\ell \hat{Y}^\ell \hat{Z}^{\rm d}}(u^\ell,w^\ell,y^\ell,z^{\rm d})[\rho(u^{\ell-d},m(z^{\rm d},w^\ell,y^\ell))+\rho_{\max}\cdot d]\\
&= \boldsymbol{E}\left\{\frac{1}{n}\sum_{i=0}^{n/\ell-1}\left[\sum_{\tau=1}^{\ell-d}\rho(u_{i\ell+\tau},f'_{\tau+d}(W_{i\ell+\tau+d},y_{i\ell+\tau+d},z^{\rm d}_{i\ell+\tau+d}))+\rho_{\max}\cdot d\right]\right\}\\
&\le \boldsymbol{E}_W\left\{\frac{1}{n}\sum_{t=1}^{n}\rho(u_t,V_t)\right\}+\frac{\rho_{\max}\cdot d}{\ell}, \qquad (55)
\end{aligned}$$

where $\boldsymbol{E}_W$ denotes expectation w.r.t. W^n only. The first inequality follows from the definition of G as the optimal estimator and the fact that, due to the delay, the estimates of the last d source symbols are not yet available in the current ℓ–block, but their distortions cannot exceed $\rho_{\max}$. The subsequent equality is by the definitions of all the ingredients involved, and by passing from empirical distributions back to time averages. Thus,

$$I(\hat{U}^\ell;\hat{Y}^\ell|W^\ell)+\log s_{\rm d} \ge \ell\cdot R^{\rm wz}_{\hat{U}^\ell|W^\ell}\left(\boldsymbol{E}_W\left\{\frac{1}{n}\sum_{i=1}^{n}\rho(u_i,V_i)\right\}+\frac{\rho_{\max}\cdot d}{\ell}\right), \qquad (56)$$

and consequently, following the second to the last line of (50), we have

$$I(\hat{U}^\ell,\hat{Y}^\ell) \quad \ge \quad I(\hat{U}^\ell;\hat{Y}^\ell|W^\ell)$$

$$
\begin{aligned}
&= I(\hat{U}^\ell;\hat{Y}^\ell|W^\ell) + \log s_{\mathrm{d}} - \log s_{\mathrm{d}} \\
&\geq \ell \cdot R^{\mathrm{WZ}}_{\hat{U}^\ell|W^\ell}\left(\boldsymbol{E}_W\left\{\frac{1}{n}\sum_{i=1}^{n}\rho(u_i,V_i)\right\} + \frac{\rho_{\max}\cdot d}{\ell}\right) - \log s_{\mathrm{d}}. \qquad (57)
\end{aligned}
$$

Finally, combining this with (49) and taking the expectation w.r.t. the randomness of the channels, we have

$$
\begin{aligned}
\ell \cdot C(\Gamma) + \log s_{\mathrm{d}} + \ell \cdot \left[\frac{s_{\mathrm{e}}\alpha^\ell \log\gamma}{\sqrt{n}} + +o\left(\frac{1}{\sqrt{n}}\right)\right] &\geq \boldsymbol{E}\left\{R^{\mathrm{WZ}}_{\hat{U}^\ell|W^\ell}\left(\boldsymbol{E}_W\left\{\frac{1}{n}\sum_{i=1}^{n}\rho(u_i,V_i)\right\} + \frac{\rho_{\max}\cdot d}{\ell}\right)\right\} \\
&\geq R^{\mathrm{WZ}}_{\hat{U}^\ell|W^\ell}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{E}\left\{\rho(u_i,V_i)\right\} + \frac{\rho_{\max}\cdot d}{\ell}\right), \qquad (58)
\end{aligned}
$$

where in the last inequality we have used Jensen's inequality and the convexity of the W-Z rate–distortion function [3, Lemma 15.9.1]. The assertion of Theorem 2 now follows immediately.

5 Variations, Modifications and Extensions

In this section, we outline a few variants of our main results that require a few changes in the model considered. We discuss the following modifications: (i) the additional constraint of common reconstruction, (ii) the case where side information is available to the encoder too, and (iii) channel state information at the encoder. In all these cases, we discuss the changes needed in proof of Theorem 2, and one can also obtain a lower bound to the excess distortion probability by applying the appropriate change of measures, following the same ideas as in the proof of Theorem 1, and as discussed after Theorem 2.

5.1 Common Reconstruction

In [23], Steinberg studied a version of the W-Z problem [24], where there is an additional constraint that the encoder would be capable of generating an *exact* copy of the reconstruction sequence to be generated by the decoder, with motivation in medical imaging, etc. In the ordinary W-Z setting, this is not the case since the reconstruction depends on the side information, which is not available to the encoder. Steinberg's solution to the W-Z problem with common reconstruction is very similar to the solution of the regular W-Z problem: the only difference is that the estimator at the decoder is allowed to be a function of the compressed representation only, rather than being a function of both the compressed representation and the side information vector. In other words, in Steinberg's scheme, the side information serves the decoder only for the purpose of binning, and not for both binning and estimation, as in the classical W-Z achievability scheme. For a pair of memoryless correlated sources, $\{(U_i, W_i)\}$, Steinberg's coding theorem [23, Theorem 1] for coding under the common reconstruction constraint, asserts that the corresponding rate–distortion function is given by

$$R_{U|W}^{\mathrm{WZ,cr}}(D) = \min I(U;V|W) \equiv \min\{I(U;V) - I(W;V)\}, \tag{59}$$

where the minimum is over all conditional distributions, $\{P_{V|U}\}$, such that $V \to U \to W$ is a Markov chain and $\boldsymbol{E}\{\rho(U,V)\} \le D$.

Equipped with this background, we can impose the common reconstruction constraint in our setting too, provided that the model of the finite–state decoder is somewhat altered: Instead of feeding the finite–state decoder sequentially by $\{(w_i, y_i)\}$, as before, we now feed it by a single sequence, $\{r_i\}$, where $r^n = (r_1, \ldots, r_n)$ is a deterministic function of u^n, which with very high probability (for large n), can be reconstructed faithfully at the decoder as a function of (w^n, y^n).

The modifications needed in the proof of Theorem 2 are in two places only: The first modification is that in the last line of eq. (50), $R_{\hat{U}^\ell|W^\ell}^{\mathrm{WZ}}[\Delta(\hat{U}^\ell|\hat{Y}^\ell, W^\ell)/\ell]$ should be replaced by $R_{\hat{U}^\ell|W^\ell}^{\mathrm{WZ,cr}}[\boldsymbol{E}\{\rho(\hat{U}^\ell, \hat{V}^\ell)\}/\ell]$, where following [23], $R_{\hat{U}^\ell|W^\ell}^{\mathrm{WZ,cr}}(D)$ is defined according to

$$R_{\hat{U}^\ell|W^\ell}^{\mathrm{WZ,cr}}(D) = \frac{1}{\ell}\min I(\hat{U}^\ell; V^\ell|W^\ell), \tag{60}$$

where the minimum is over all $\{P_{\hat{V}^\ell|\hat{U}^\ell}\}$ such that $\hat{V}^\ell \to \hat{U}^\ell \to W^\ell$ is a Markov chain and $\boldsymbol{E}\{\rho(\hat{U}^\ell, \hat{V}^\ell) \le \ell \cdot D$. The second modification is in eq. (55), where $G(w^\ell, y^\ell, z^{\mathrm{d}})$, $m(z^{\mathrm{d}}, w^\ell, y^\ell)$ and $f_{\tau+d}(W_{i\ell+\tau+d}, y_{i\ell+\tau+d}, z^{\mathrm{d}}_{i\ell+\tau+d})$ should be replaced by $G(r^\ell, z^{\mathrm{d}})$, $m(z^{\mathrm{d}}, r^\ell)$, and $f_{\tau+d}(r_{i\ell+\tau+d}, z^{\mathrm{d}}_{i\ell+\tau+d})$, respectively.

The achievability is based on source coding using Steinberg's coding scheme [23], followed by a capacity–achieving channel code.

5.2 Side Information at Both Ends

So far, we have considered the case where the side information is available at the decoder only. On the face of it, one might argue that there is no much point to address the case where side information is available to both encoder and decoder, because it is much easier and it can even be viewed as a special case of side information at the decoder only (simply by redefining $\{(u_i, w_i)\}$ as the "source"). Nevertheless, we mention the case of two–sided side information for two reasons:

1. We can allow $\boldsymbol{w}$ to be an individual sequence too, in addition to $\boldsymbol{u}$, as opposed to our assumption so far that it is generated by a DMC fed by $\boldsymbol{u}$.

2. We can derive more explicit lower bounds to the distortion. These bounds automatically apply also to the case where the side information is available to the decoder only, although they might not be tight for that case.

In the purely probabilistic setting, the rate–distortion function in the presence of side information at both ends is given by the so called conditional rate–distortion function. As discussed in [25], the only difference between the W-Z rate–distortion function and the conditional rate–distortion function is that in the former, there is the constraint of the Markov structure, whereas the conditional rate–distortion function this constraint is dropped. The proof of Theorem 2 can easily be altered to incorporate two–sided availability of side information, with both $\boldsymbol{u}$ and $\boldsymbol{w}$ being deterministic sequences. The only modification needed is in eq. (50), which will now read as follows:

$$I(\hat{U}^\ell; \hat{Y}^\ell) \geq \ell \cdot R_{\hat{U}^\ell}[\Delta(\hat{U}^\ell|\hat{Y}^\ell)/\ell] \geq \ell \cdot R_{\hat{U}^\ell|\hat{W}^\ell}[\Delta(\hat{U}^\ell|\hat{Y}^\ell, \hat{W}^\ell)/\ell], \tag{61}$$

where the second inequality is due to the fact that ignoring side information at both ends cannot be better than using it optimally. Note that we have also replaced W^ℓ by $\hat{W}^\ell$, to account for the fact that we allow it to be a deterministic sequence too, as mentioned before. Obviously, achievability is by conditional rate–distortion coding followed by capacity–achieving channel coding.

For the purpose of the lower bound to the distortion, we can further lower bound the empirical conditional rate–distortion function as follows in the spirit of the conditional Shannon lower bound [9]. First, we can represent it as

$$R_{\hat{U}^\ell|\hat{W}^\ell}(D) = H(\hat{U}^\ell|\hat{W}^\ell) - \max_{\{P_{\hat{V}^\ell|\hat{U}^\ell\hat{W}^\ell}:\ \boldsymbol{E}\{\rho(\hat{U}^\ell,\hat{V}^\ell)\leq \ell D\}} H(\hat{U}^\ell|\hat{W}^\ell, \hat{V}^\ell). \tag{62}$$

Now, the first term, $H(\hat{U}^\ell|\hat{W}^\ell)$, can be further lower bounded (within an asymptotically negligible term) in terms of the conditional Lempel-Ziv code–length function of u^n given w^n (as side information at both ends), as defined in [29]. Specifically, the following inequality is derived in [15, eq.

(17)]:

$$H(\hat{U}^\ell|\hat{W}^\ell) \geq \frac{\ell}{n}\sum_{j=1}^{c(w^n)}[c_j(u^n|w^n)+q^2]\log\frac{c_j(u^n|w^n)}{4q^2}, \tag{63}$$

where q is a constant that depends only on ℓ and on the sizes of $\mathcal{U}$ and $\mathcal{W}$, $c(w^n)$ denotes the number of distinct phrases of w^n that appear in joint incremental parsing [30] of (u^n, w^n), and $c_j(u^n|w^n)$ is the number of distinct phrases of u^n that appear jointly with the j–th distinct phrase of w^n. Similarly as in [16, Theorem 2], for the case where $\mathcal{U} = \mathcal{V} = \{0, 1, \ldots, \alpha - 1\}$ and a difference distortion measure, $\rho(u, v) = \varrho(u - v)$ (where the subtraction is defined modulo α), the second, subtracted term of (62), can be easily upper bounded by the constrained maximum entropy function, i.e.,

$$\max_{\{P_{\hat{V}^\ell|\hat{U}^\ell\hat{W}^\ell}:\ \boldsymbol{E}\{\rho(\hat{U}^\ell,\hat{V}^\ell)\leq \ell D\}} H(\hat{U}^\ell|\hat{W}^\ell,\hat{V}^\ell) \leq \ell\cdot\Phi(D), \tag{64}$$

where

$$\Phi(D) = \sup_{\theta\geq 0}\left[\theta D + \log\left(\sum_{u\in\mathcal{U}}2^{-\theta\varrho(u)}\right)\right], \tag{65}$$

as can easily be shown by the standard solution to the problem of maximum entropy under a moment constraint. It then follows that the expected distortion is lower bounded in terms of the inverse function, $\Psi = \Phi^{-1}$, computed at the difference,

$$\frac{1}{n}\sum_{j=1}^{c(w^n)}c_j(u^n|w^n)\log c_j(u^n|w^n) - C(\Gamma) - \eta(s_{\mathrm{e}}, s_{\mathrm{e}}, d, \ell, n)$$

where $\eta(s_{\mathrm{e}}, s_{\mathrm{d}}, d, \ell, n)$ accounts for all resulting redundancy terms (similarly as in [16, proof of Theorem 2]). Specifically, the inverse function, Ψ, is given by (see Appendix B)

$$\Psi(R) = \inf_{\vartheta\geq 0}\vartheta\cdot\left[R - \log\left(\sum_{u\in\mathcal{U}}2^{-\varrho(u)/\vartheta}\right)\right], \tag{66}$$

and so,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{E}\{\rho(u_i,V_i)\} \quad \geq \quad & \sup_{\vartheta\geq 0}\vartheta\cdot\left[\frac{1}{n}\sum_{j=1}^{c(w^n)}c_j(u^n|w^n)\log c_j(u^n|w^n) - C(\Gamma) - \right.\\ & \left.\eta(s_{\mathrm{e}}, s_{\mathrm{d}}, d, \ell, n) - \log\left(\sum_{u\in\mathcal{U}}2^{-\varrho(u)/\vartheta}\right)\right] - \frac{\rho_{\max}d}{\ell}.\end{aligned} \tag{67}$$

Of course, instead of maximizing over ϑ, one may select any arbitrary positive ϑ and thereby obtain a a valid lower bound, albeit not as tight. We observe that whenever the conditional LZ complexity,

$\frac{1}{n}\sum_{j=1}^{c(w^n)} c_j(u^n|w^n)\log c_j(u^n|w^n)$, exceeds the channel capacity (plus the redundancy terms), the expected distortion has a non–trivial, strictly positive lower bound.

The advantage of the use of the conditional LZ complexity is that it is easier to calculate than the ℓ–th order empirical entropy, especially when ℓ is large, as the super-alphabet size grows exponentially with ℓ. In other words, we sacrifice tightness to a certain extent, at the benefit of facilitating the calculation of the bound.

5.3 Channel–State Information at the Encoder

The last extension that we discuss in this work is from the DMC, $P_{Y|X}$, to a state–dependent channel model, $P_{Y^n|X^nS^n} = [P_{Y|XS}]^n$, where the state sequence, s^n, which is governed by a discrete memoryless source, $P_{S^n} = [P_S]^n$ of alphabet $\mathcal{S}$ of size σ, is fed into the encoder as well. Specifically, the finite–state encoder is now described by the following set of recursive equations:

$$\begin{aligned} t &= i \bmod \ell && (68)\\ x_i &= f_t(u_i, s_i, z_i^{\mathrm{e}}), && (69)\\ z_{i+1}^{\mathrm{e}} &= g_t(u_i, s_i, z_i^{\mathrm{e}}). && (70)\end{aligned}$$

Since $\{s_i\}$ are fed sequentially into the encoder, this clearly falls in the category of causal state information [21]. Here, eq. (46) in the proof of Theorem 2 should be modified as follows. Let $P_{\hat{U}^\ell S^\ell} = P_{\hat{U}^\ell} \times P_{S^\ell} = P_{\hat{U}^\ell} \times [P_S]^\ell$. We also define

$$\mathcal{P}_\ell(\Gamma) = \{(P_B, L):\ X^\ell = L(B, S^\ell), \boldsymbol{E}_{P_{X^\ell}}\phi(X^\ell) \le \ell\Gamma\}, \tag{71}$$

where B is an auxiliary random variable whose alphabet size need not be larger than $\min\{\beta^\ell - 1, \sigma^\ell + 1, \gamma^\ell\}$, and is independent of S^ℓ [6, Theorem 7.2], and where $\phi(\hat{X}^\ell)$ is the additive extension of the single–letter transmission cost function, that is, $\phi(\hat{X}^\ell) = \sum_{i=1}^\ell \phi(\hat{X}_i)$. Then,

$$\begin{aligned} \boldsymbol{E}\{I(\hat{U}^\ell; \hat{Y}^\ell)\} &= \boldsymbol{E}\{H(\hat{Y}^\ell)\} - \boldsymbol{E}\{H(\hat{Y}^\ell|\hat{U}^\ell)\} && (72)\\ &\le H(Y^\ell) - \boldsymbol{E}\{H(\hat{Y}^\ell|\hat{U}^\ell)\} && (73)\\ &\le H(Y^\ell) - H(Y^\ell|\hat{U}^\ell) + \ell\cdot o(n) && (74)\\ &= I(\hat{U}^\ell; Y^\ell) + \ell\cdot o(n) && (75)\\ &\le \max_{(P_B,L)\in\mathcal{P}_\ell(\Gamma)} I(B; Y^\ell) + \ell\cdot o(n) && (76)\end{aligned}$$

$$\leq \quad \ell \cdot [C_{\rm S}(\Gamma) + o(n)]\,, \tag{77}$$

where $C_{\rm S}(\Gamma)$ is the capacity of the Shannon channel of causal channel state information [21] with average transmission cost limited by Γ [10, eqs. (3.9), (3.33)]. The four inequalities of this chain are explained as follows. The first inequality is due to the concavity of the entropy as a functional of the underlying distribution. The second one is obtained by the weak law of large numbers, which guarantees that $P_{\hat{Y}^\ell|X^\ell S^\ell} \to P_{Y^\ell|X^\ell S^\ell}$, in probability, as $n \to \infty$ (for fixed ℓ), and therefore,

$$P_{\hat{Y}^\ell|\hat{U}^\ell}(y^\ell|u^\ell) = \sum_{s^\ell, z^{\rm e}} P_{\hat{Z}^{\rm e}|\hat{U}^\ell}(z^{\rm e}|u^\ell) P_{S^\ell}(s^\ell) P_{\hat{Y}^\ell|X^\ell S^\ell}(y^\ell|q[u^\ell, s^\ell, z^{\rm e}], s^\ell)$$

tends to $P_{Y^\ell|\hat{U}^\ell}(y^\ell|u^\ell)$ in probability for all $u^\ell \in \mathcal{U}^\ell$, $y^\ell \in \mathcal{Y}^\ell$. The third inequality is due to the fact that $\hat{U}^\ell$ (like any feasible B) is independent of S^ℓ and since the reference encoding function, q, is assumed to comply with the transmission cost constraint. Note that the additional dependence of x^ℓ upon $z^{\rm e}$ can be viewed as randomized encoding according to

$$P(x^\ell|u^\ell, s^\ell) = \sum_{z^{\rm e}} P_{\hat{Z}^{\rm e}|\hat{U}^\ell}(z^{\rm e}|u^\ell) \cdot 1\{x^\ell = q[u^\ell, s^\ell, z^{\rm e}]\},$$

that cannot improve capacity. Finally, the last inequality is due to the fact that the multi-letter extension of the capacity formula cannot improve on the single–letter version, as can easily been seen by comparing the highest rate achievable by multi–letter random coding (over the superalphabet of ℓ–vectors) to the converse bound on the highest achievable rate, which is given by the single–letter formula.

Finally, note that if the allowed delay, d, exceeds the period ℓ, and $s_{\rm e} \geq \alpha^\ell$, then the encoder can afford to wait until the end of the ℓ–block (and store it as its state) before beginning to encode. In this case, the more relevant capacity formula is the one of non–causal state information [8], which is, in general, larger than $C_S(\Gamma)$, and hence serves as an upper bound as well.

Appendix A

In this appendix, we prove eq. (47). The proof is very similar to the proof of eq. (32) in [16], but here we derive a tighter redundancy term by exploiting the fact that the number of distinct pairs $\{(u^\ell, x^\ell)\}$, that may appear as non–overlapping ℓ-blocks of (u^n, x^n), cannot really be as large as $(\alpha\beta)^\ell$, but only $s_{\rm e} \cdot \alpha^\ell$ at most. The simple reason is that x^ℓ is a deterministic function of $(z^{\rm e}, u^\ell)$, which in turn takes on at most $s_{\mathrm{e}} \cdot \alpha^\ell$ different values. Although the proof is very similar to that of [16], we provide it here in full detail, for the sake of completeness.

As explained in [16], we invoke the following result (see [1], [2] and [19, Proposition 5.2] therein, as well as [18, Appendix A]): Let $\hat{P}_n$ be the first order empirical distribution associated with an n–sequence drawn from a memoryless m–ary source P with alphabet $\{1,\ldots,m\}$. Then,

$$n \cdot \boldsymbol{E}\{D(\hat{P}_n \| P)\} = \frac{(m-1)\log e}{2} + o(1), \tag{A.1}$$

which is equivalent to

$$\boldsymbol{E}\left\{-\sum_{k=1}^{m} \hat{P}_n(k) \log \hat{P}_n(k)\right\} = H - \frac{(m-1)\log e}{2n} - o\left(\frac{1}{n}\right), \tag{A.2}$$

where H is the entropy of P. We now apply this result to the 'source' $P(y^\ell|u^\ell, z^{\mathrm{e}}) \equiv P(y^\ell|x^\ell)$ for every pair (u^ℓ, z^{e}) that appears more than $\epsilon n/\ell$ times as ℓ–blocks along the (deterministic) sequence pair (u^n, x^n).

$$\begin{aligned}
& \boldsymbol{E} H(\hat{Y}^\ell|\hat{U}^\ell, \hat{Z}^{\mathrm{e}}) \\
= \; & \boldsymbol{E}\left\{\sum_{u^\ell, z^{\mathrm{e}}} P_{\hat{U}^\ell \hat{Z}^{\mathrm{e}}}(u^\ell, z^{\mathrm{e}}) H(\hat{Y}^\ell|\hat{U}^\ell = u^\ell, \hat{Z}^{\mathrm{e}} = z^{\mathrm{e}})\right\} \\
\ge \; & \sum_{\{u^\ell, z^{\mathrm{e}}:\; P_{\hat{U}^\ell \hat{Z}^{\mathrm{e}}}(u^\ell, z^{\mathrm{e}}) \ge \epsilon\}} P_{\hat{U}^\ell \hat{Z}^{\mathrm{e}}}(u^\ell, z^{\mathrm{e}}) \cdot \boldsymbol{E}\{H(\hat{Y}^\ell|\hat{U}^\ell = u^\ell, \hat{Z}^{\mathrm{e}} = z^{\mathrm{e}})\} \\
= \; & \sum_{\{u^\ell, z^{\mathrm{e}}:\; P_{\hat{U}^\ell \hat{Z}^{\mathrm{e}}}(u^\ell, z^{\mathrm{e}}) \ge \epsilon\}} P_{\hat{U}^\ell \hat{Z}^{\mathrm{e}}}(u^\ell, z^{\mathrm{e}}) \left[H(Y^\ell|\hat{X}^\ell = q(z^{\mathrm{e}}, u^\ell)) - \frac{(\gamma^\ell - 1)\log e}{2nP_{\hat{U}^\ell \hat{Z}^{\mathrm{e}}}(u^\ell, z^{\mathrm{e}})/\ell} - o\left(\frac{\ell}{n\epsilon}\right)\right] \\
\ge \; & \sum_{\{u^\ell, z^{\mathrm{e}}:\; P_{\hat{U}^\ell \hat{Z}^{\mathrm{e}}}(u^\ell, z^{\mathrm{e}}) \ge \epsilon\}} P_{\hat{U}^\ell \hat{Z}^{\mathrm{e}}}(u^\ell, z^{\mathrm{e}}) H(Y^\ell \| \hat{X}^\ell = q(z^{\mathrm{e}}, u^\ell)) - \frac{\ell s_{\mathrm{e}}(\alpha\gamma)^\ell \log e}{2n} - o\left(\frac{\ell}{n\epsilon}\right) \\
= \; & \sum_{u^\ell, z^{\mathrm{e}}} P_{\hat{U}^\ell \hat{Z}^{\mathrm{e}}}(u^\ell, z^{\mathrm{e}}) H(Y^\ell|\hat{X}^\ell = q(z^{\mathrm{e}}, u^\ell)) - \\
& \sum_{\{u^\ell, x^\ell:\; P_{\hat{U}^\ell \hat{X}^\ell}(u^\ell, x^\ell) < \epsilon\}} P_{\hat{U}^\ell \hat{Z}^{\mathrm{e}}}(u^\ell, z^{\mathrm{e}}) H(Y^\ell|\hat{X}^\ell = q(z^{\mathrm{e}}, u^\ell)) - \frac{\ell s_{\mathrm{e}}(\alpha\gamma)^\ell \log e}{2n} - o\left(\frac{\ell}{n\epsilon}\right) \\
\ge \; & H(Y^\ell|\hat{X}^\ell) - \epsilon s_{\mathrm{e}} \alpha^\ell \cdot \ell \log \gamma - \frac{\ell s_{\mathrm{e}}(\alpha\gamma)^\ell \log e}{2n} - o\left(\frac{\ell}{n\epsilon}\right) \\
\stackrel{\Delta}{=} \; & H(Y^\ell|\hat{X}^\ell) - \ell \cdot \Delta_0(s_{\mathrm{e}}, \epsilon, \ell, n).
\end{aligned} \tag{A.3}$$

Selecting $\epsilon = 1/\sqrt{n}$, we have

$$\Delta_1(s_{\mathrm{e}}, \ell, n) = \Delta_0\left(s_{\mathrm{e}}, \frac{1}{\sqrt{n}}, \ell, n\right) = \frac{s_{\mathrm{e}} \alpha^\ell \log \gamma}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right), \tag{A.4}$$

as defined in eq. (48). This completes the proof of eq. (47).

Appendix B

In this appendix, we prove that the function $\Psi(\cdot)$, defined in (66), is the inverse of the function $\Phi(D)$, defined in (65). Let us denote

$$R = \Phi(D) = \sup_{\theta \geq 0} \left[\theta D + \log \left(\sum_{u \in \mathcal{U}} 2^{-\theta \varrho(u)} \right) \right]. \tag{B.1}$$

This means that:

1. $\forall\, \theta \geq 0$,
$$R \geq \theta D + \log \left(\sum_{u \in \mathcal{U}} 2^{-\theta \varrho(u)} \right). \tag{B.2}$$

2. There exists a positive sequence, $\{\theta_n\}$, such that
$$\lim_{n \to \infty} \left[\theta_n D + \log \left(\sum_{u \in \mathcal{U}} 2^{-\theta_n \varrho(u)} \right) \right] = R. \tag{B.3}$$

But this is clearly equivalent to the set of statements:

1. $\forall\, \theta \geq 0$,
$$D \leq \frac{R - \log \left(\sum_{u \in \mathcal{U}} 2^{-\theta \varrho(u)} \right)}{\theta}. \tag{B.4}$$

2. $\exists\, \theta \geq 0$,
$$\lim_{n \to \infty} \frac{R - \log \left(\sum_{u \in \mathcal{U}} 2^{-\theta_n \varrho(u)} \right)}{\theta_n} = D. \tag{B.5}$$

This in turn is equivalent to the statement that

$$D = \inf_{\theta \geq 0} \frac{R - \log \left(\sum_{u \in \mathcal{U}} 2^{-\theta \varrho(u)} \right)}{\theta} = \inf_{\vartheta \geq 0} \vartheta \cdot \left[R - \log \left(\sum_{u \in \mathcal{U}} 2^{-\varrho(u)/\vartheta} \right) \right] = \Psi(R). \tag{B.6}$$

References

[1] K. Atteson, "The asymptotic redundancy of Bayes rules for Markov chains," *IEEE Trans. Inform. Theory*, vol. 45, no. 6, pp. 2104–2109, September 1999.

[2] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, pp. 453-471, May 1990.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Hoboken, New Jersey, 2006.

[4] I. Csiszár, "Joint source–channel error exponent," *Problems of Control and Information Theory*, vol. 9, no. 5, pp. 315–328, 1980.

[5] I. Csiszár, "On the error exponent of source-channel transmission with a distortion threshold," *IEEE Trans. Inform. Theory*, vol. IT–28, no. 6, pp. 823–828, November 1982.

[6] A. El Gamal and Y.-H. Kim, *Network Information Theory*, Cambridge University Press, Cambridge, 2011.

[7] R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley & Sons, New York, 1968.

[8] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Problems of Information and Control*, vol. 9, no. 1, pp. 19-31, 1980.

[9] R. M. Gray, "A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions," *IEEE Trans. inform. Theory*, vol. IT-19, no. 4, pp. 480–489, July 1973.

[10] G. Keshet, Y. Steinberg, and N. Merhav, "Channel coding in the presence of side information," *Foundations and Trends in Communications and Information Theory*, vol. 4, no. 6, pp. 445–586, 2007.

[11] A. Lapidoth, A. Malär, and M. Wigger, "Constrained source–coding with side information," *IEEE Trans. Inform. Theory*, vol. 66, no. 6, pp. 3218–3237, June 2014.

[12] A. Lempel and J. Ziv, "On the complexity of finite sequences," *IEEE Trans. Inform. Theory*, vol. IT–22, no. 1, pp. 75–81, January 1976.

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

[14] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT–20, no. 2, pp. 197–199, March 1974.

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

[16] N. Merhav, "On the data processing theorem in the semi–deterministic setting," *IEEE Trans. Inform. Theory*, vol. 60, no. 10, pp. 6032–6040, October 2014.

[17] N. Merhav and S. Shamai (Shitz), "On joint source–channel coding for the Wyner–Ziv source and the Gel'fand–Pinsker channel," *IEEE Trans. Inform. Theory*, vol. 49, no. 11, pp. 2844–2855, November 2003.

[18] N. Merhav and M. J. Weinberger, "On universal simulation of information sources using training data," *IEEE Trans. Inform. Theory*, vol. 50, no. 1, pp. 5–20, January 2004.

[19] N. Merhav and J. Ziv, "On the Wyner–Ziv problem for individual sequences," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 867–873, March 2006.

[20] S. Shamai (Shitz), S. Verdú and R. Zamir, "Systematic lossy source/ channel coding," *IEEE Trans. Inform. Theory*, vol. 44, no. 2, pp. 564–579, March 1998.

[21] C. E. Shannon, "Channels with side information at the transmitter," *IBM Journal Research and Development*, vol. 2, pp. 289-293, October 1958.

[22] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding in discrete memoryless channels," *Inform. Contr.*, vol. 10, pp. 65–103 and 522–552, 1967.

[23] Y. Steinberg, "Coding and common reconstruction," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 4995-–5010, November 2009.

[24] 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.

[25] R. Zamir, "The rate loss in the Wyner–Ziv problem," *IEEE Trans. Inform. Theory*, vol. 42, no. 6, pp. 2073–2084, November 1996.

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

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

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

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

[30] 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.