*IRWIN AND JOAN JACOBS*
*CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES*

# On the Data Processing Theorem in the Semi-Deterministic Setting

**Neri Merhav**

**CCIT Report #828**
**March 2013**


*DEPARTMENT OF ELECTRICAL ENGINEERING*
*TECHNION - ISRAEL INSTITUTE OF TECHNOLOGY, HAIFA 32000, ISRAEL*

CCIT Report #828 March 2013

# On the Data Processing Theorem in the Semi–Deterministic Setting

Neri Merhav

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

**Abstract**

Data processing lower bounds on the expected distortion are derived in the finite–alphabet semi–deterministic setting, where the source produces a deterministic, individual sequence, but the channel model is probabilistic, and the decoder is subjected to various kinds of limitations, e.g., decoders implementable by finite–state machines, with or without counters, and with or without a restriction of common reconstruction with high probability. Some of our bounds are given in terms of the Lempel–Ziv complexity of the source sequence or the reproduction sequence. We also demonstrate how some analogous results can be obtained for classes of linear encoders and linear decoders in the continuous alphabet case.

**Index Terms:** Data processing theorem, finite–state machine, Lempel–Ziv algorithm, redundancy, delay, common reconstruction.

## 1 Introduction

In a series of articles from the seventies and the eighties of the twentieth century, Ziv [10],[11],[12], and Ziv and Lempel [3], [13], have created a theory of universal source coding for individual sequences using finite–state machines. In particular, the work [10] focuses on universal, fixed–rate, (almost) lossless compression of individual sequences using finite–state encoders and decoders, which was then further developed to the famous Lempel–Ziv algorithm [3], [13]. In [11], the framework of [10] was extended to lossy coding for both noiseless and noisy transmission (subsections II.A and II.B of [11], respectively), and later further extended in other directions, such as incorporation of side information in the context of almost lossless compression, where the side information data

is also modeled as an individual sequence [12], in other words, an individual–sequence counterpart of Slepian–Wolf coding [8] was studied in [12] (see also a later extension to the lossy case [7]).

The main trigger for this paper stems from the coding theorem for noisy transmission in [11, Subsection II.B]. We begin by revisiting the assertion and the proof of the converse part of this theorem (Theorem 3 and eqs. (12) and (13) in [11]), which provides a lower bound on the distortion in a *semi–deterministic setting*, where the source emits a deterministic (individual) sequence, but the channel model is probabilistic as usual (in particular, it is a discrete memoryless channel) and the encoder and decoder are limited to be finite–state machines with no more than $s$ states and a given overall delay, which we shall denote by $d$. While this theorem is essentially correct, it turns out that there are certain imprecise steps in its proof (see Appendix for details) and moreover, in relation to our corrections to this proof, the assertion of the theorem itself can be strengthened and sharpened. The revisited converse theorem imposes no limitations on the encoder,[1] and allows the decoder to be equipped with a modulo–$\ell$ counter ($\ell$ – positive integer) in addition to its $s$ states of memory, which means that within each period of length $\ell$, the decoder is allowed to be time–varying, as opposed to the time–invariant model used in [11] and in related papers.[2] Also, our lower bound on the distortion depends, not only on the number of states $s$ (as in [11]), but also on the allowed delay $d$ (as well as on some additional redundancy terms).

Beyond the above described revisit of Theorem 3 of [11], we also derive additional lower bounds on the expected distortion in the semi–deterministic setting. One of them is associated with a restriction of a *common reconstruction* (with high probability) at both encoder and decoder, which is a setup that has recently received some attention in other contexts, like the Wyner–Ziv problem (see e.g., [9]), with motivations in medical imaging, etc. In addition, some of our bounds are given in a more explicit form, in terms of the Lempel–Ziv complexity of the source sequence or the reproduction sequence. This may be interesting in the sense that the Lempel–Ziv complexity usually arises when the finite–state structure is imposed on the encoder, whereas in our case, it is imposed on the decoder. Finally, we demonstrate how some analogous results can be obtained for

[1] The assumption that the encoder is a finite–state machine is not really used in [11] either,

[2] One might argue that a finite–state machine with $s$ states and a modulo–$\ell$ counter is just a particular finite–state machine with a total number of $s \cdot \ell$ states. While this argument is true, in principle, the idea is that this partition of the total number of allowed states between those that are allocated to implement a clock (the counter) and those that are allocated to memory of past input data (the remaining $s$ states) give us more detailed and more refined results.

classes of linear encoders and linear decoders in the continuous alphabet case.

It should be emphasized that our focus in this paper is primarily on lower bounds and converse theorems, and not quite on achievability schemes. Most of our bounds can be asymptotically approached by conceptually simple, separation–based schemes, in the spirit of the one proposed in [11] or with certain modifications and variations on the same ideas.

The outline of this paper is as follows. In Section 2, we establish notation conventions and formalize the semi–deterministic setting under consideration. In Section 3, we derive a lower bound on the distortion without the common reconstruction requirement, and in Section 4, we derive the parallel lower bound under common reconstruction. In both sections, we also derive the aforementioned alternative lower bounds, which can be calculated more easily. Finally, in Section 5, we give an outline of an analogue of the main result of Section 2 for continuous alphabets and linear encoders and decoders.

## 2 Problem Formulation and 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. 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 $\boldsymbol{u} = (u_1, u_2, \ldots)$ be an individual source sequence of symbols in a finite alphabet $\mathcal{U}$ of cardinality $|\mathcal{U}| = J$. The sequence $\boldsymbol{u}$ is encoded using a general encoder, whose output at time $t$ is $x_t \in \mathcal{X}$, where $\mathcal{X}$ is another finite alphabet[3] of size $|\mathcal{X}| = K$. The sequence $\boldsymbol{x} = (x_1, x_2, \ldots)$ is fed

[3]In the general formulations of the joint source–channel coding problem, the source and the channel are allowed to operate at different rates, and then, in the case of block codes, source blocks of a given length may be mapped into channel blocks of a different length. This degree of freedom, however, is essentially available here too, by redefining $\mathcal{U}$ and $\mathcal{X}$ to be superalphabets of the appropriate sizes.

into a discrete memoryless channel (DMC), characterized by the matrix of single-letter transition probabilities $\{P(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$, where the output alphabet $\mathcal{Y}$ is a finite alphabet of size $|\mathcal{Y}| = L$. The channel output $\boldsymbol{y} = (y_1, y_2, \ldots)$ is in turn fed into a finite–state decoder, which is defined by the following recursive equations:

$$v_{t-d} = f(z_t, y_t), \qquad t = d+1, d+2, \ldots \tag{1}$$

$$z_{t+1} = g(z_t, y_t), \qquad t = 1, 2, \ldots \tag{2}$$

where $z_t \in \mathcal{Z}$ is the decoder state at time $t$, $\mathcal{Z}$ being a finite set of states of size $s$, $v_{t-d} \in \mathcal{V}$ is the reconstructed sequence, delayed by $d$ time units ($d$ – positive integer) and $f : \mathcal{Z} \times \mathcal{Y} \to \mathcal{V}$ and $g : \mathcal{Z} \times \mathcal{Y} \to \mathcal{Z}$ are the output function and the next–state function, respectively. The reconstruction alphabet $\mathcal{V}$ of size $M$.

A slightly more sophisticated model allows the decoder to be equipped with a modulo–$\ell$ counter, in addition to its state variable. This means that the functions $f$ and $g$ are allowed to be time–varying within each period of length $\ell$. In particular, in this case, the decoding equations would admit the form:

$$\tau = t \bmod \ell, \qquad t = 1, 2, \ldots \tag{3}$$

$$v_{t-d} = f_\tau(z_t, y_t), \qquad t = d+1, d+2, \ldots \tag{4}$$

$$z_{t+1} = g_\tau(z_t, y_t), \qquad t = 1, 2, \ldots \tag{5}$$

In some applications, one may be interested in a common reconstruction at both the encoder and decoder (with high probability). In our context, this means that for a certain positive integer, which we will choose to be $\ell$, there is a deterministic function $q : \mathcal{U}^\ell \to \mathcal{V}^\ell$ such that

$$\lim_{n\to\infty} \frac{\ell}{n} \sum_{i=0}^{n/\ell-1} \Pr\{V_{i\ell+1}^{i\ell+\ell} \neq q(u_{i\ell+1}^{i\ell+\ell})\} = 0, \tag{6}$$

where here and throughout the sequel, probabilities and expectations are defined with respect to (w.r.t.) the randomness of the channel. This means that there is a target reconstruction $\hat{v}^n$, obtained by $n/\ell$ successive applications of $q(\cdot)$, i.e., $\hat{v}_{i\ell+1}^{i\ell+\ell} = q(u_{i\ell+1}^{i\ell+\ell})$, $i = 0, 1, 2, \ldots, n/\ell - 1$, such that $V^n$ is very close to $\hat{v}^n$ in the sense of eq. (6). For example, in the traditional coding theorem of joint source–channel coding, this is achieved by separate source– and channel coding, where $\hat{v}_{i\ell+1}^{i\ell+\ell}$ are rate–distortion reproduction codewords of $u_{i\ell+1}^{i\ell+\ell}$, respectively.

For a given distortion measure $\rho : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$, we are interested in deriving lower bounds on the minimum achievable expected distortion, $\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{E}\{\rho(u_t, V_t)\}$, as functions of the alphabet sizes, the number of stated $s$, the allowed delay $d$, and the period $\ell$, if applicable, with/without a modulo–$\ell$ counter at the decoder, and with/without the requirement of common reconstruction with high probability.

Throughout our assertions and derivations, we will make heavy use of the following additional notation. Assume, without essential loss of generality, that $\ell$ divide $n$ and consider the segmentation of each $n$–vector to $n/\ell$ non–overlapping blocks of length $\ell$, that is,

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

and similar definitions for $x^n$, $y^n$, and $v^n$, 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

$$\hat{P}_{U^\ell X^\ell Y^\ell V^\ell Z}(u^\ell, x^\ell, y^\ell, v^\ell, z) = \frac{\ell}{n}\sum_{i=0}^{n/\ell-1} \mathcal{I}(\boldsymbol{u}_i = u^\ell, \boldsymbol{x}_i = x^\ell, \boldsymbol{y}_i = y^\ell, \boldsymbol{v}_i = v^\ell, z_{i\ell+1} = z), \tag{7}$$

where $\mathcal{I}(\cdot)$ is the indicator function of an event. Correspondingly, unless specified otherwise, $U^\ell$, $X^\ell$, $Y^\ell$, $V^\ell$ and $Z$ are understood to be random variables jointly distributed according to $\hat{P}_{U^\ell X^\ell Y^\ell V^\ell Z}$ and all information measures associated with them will be denoted as in the customary notation conventions of the information theory literature, but with "hats", for example, $\hat{H}(U^\ell)$ is the empirical entropy associated with $U^\ell$, $\hat{I}(X^\ell; Y^\ell)$ is the empirical mutual information between $X^\ell$ and $Y^\ell$, and so on. Accordingly, the $\ell$–th order empirical rate distortion function, associated with $u^n$ and distortion measure $\rho$, is defined as

$$\hat{R}_{U^\ell}(D) = \min\left\{\frac{1}{\ell}\hat{I}(U^\ell; \tilde{V}^\ell) : \ \boldsymbol{E}\rho(U^\ell; \tilde{V}^\ell) \le D\right\}, \tag{8}$$

where $\tilde{V}^\ell$ is a generic random variable (not to be confused with $V^\ell$, which is defined empirically), taking on values in $\mathcal{V}^\ell$, the mutual information $\hat{I}(U^\ell; \tilde{V}^\ell)$ and expected distortion $\boldsymbol{E}\rho(U^\ell, \tilde{V}^\ell)$ are defined w.r.t. $\hat{P}_{U^\ell} P_{\tilde{V}^\ell|U^\ell}$, and the minimization is across all conditional distributions $P_{\tilde{V}^\ell|U^\ell}$. Here, $\rho(U^\ell, \tilde{V}^\ell)$ is defined additively over the corresponding components of both vectors. Similarly, $\hat{D}_{U^\ell}(R)$ is the corresponding distortion–rate function, which is the inverse of $\hat{R}_{U^\ell}(D)$, and which is defined as

$$\hat{D}_{U^\ell}(R) = \min\left\{\frac{1}{\ell}\boldsymbol{E}\rho(U^\ell, \tilde{V}^\ell) : \ \hat{I}(U^\ell; \tilde{V}^\ell) \le \ell R\right\}. \tag{9}$$

In the sequel, we will define some additional empirical rate–distortion functions and distortion–rate functions, with certain modifications of the above definitions.

## 3 Distortion Bounds Without Common Reconstruction

We begin from the simpler case where there is no requirement of common reconstruction. Our first result is the following:

**Theorem 1** *Consider the communication setting described in Section 2. Let $u^n$ be an individual sequence, let $C$ be the capacity of the discrete memoryless channel, and let the overall coding–decoding delay be $d$. Then, for every decoder with $s$ states and a modulo–$\ell$ counter,*

$$\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{E}\{\rho(u_t, V_t)\} \geq \hat{D}_{U^\ell}\left(C + \frac{2\log s + d\log M}{\ell} + \delta_1(\ell, n)\right), \tag{10}$$

*where*

$$\delta_1(\ell, n) = \frac{(JK)^\ell \log L}{\sqrt{n}} + \frac{(JKL)^\ell \log e}{2n} + o\left(\frac{1}{\sqrt{n}}\right). \tag{11}$$

The interesting term, in the argument of the function $\hat{D}_{U^\ell}(\cdot)$, is the second one, namely, the term $(2\log s + d\log M)/\ell$, which seemingly plays a role of an effective "extra capacity" contributed by the state variable, that carries memory of past data from block to block and by the allowed delay. This happens because the lower bound holds for every individual sequence $u^n$ and every encoder and decoder in the allowed class, including ones that happen to be 'tailored' to $u^n$ in a certain sense (for example, the finite–state machine at the decoder may be designed to periodically produce a certain pattern that happens to be repetitive in $u^n$). The dependence on $\ell$ is much more complicated, because $\ell$ appears also in the additional term $\delta_1(\ell, n)$, and more importantly, in the function $\hat{D}_{U^\ell}(\cdot)$ itself. The lower bound is not necessarily a monotonically decreasing function of $\ell$, but this should not be surprising since the real optimum performance need not have such a monotonicity property either. For example, if $u^n$ happens to be periodic (or almost periodic) with period $\ell$, it seems plausible that it will be reproduced better by a decoder with a modulo–$\ell$ counter than by one with a modulo–$(\ell+1)$ counter, which obviously cannot keep the synchronization with $u^n$. In the absence of a modulo–$\ell$ counter at the decoder, Theorem 1 still applies, but then $\ell$ becomes just a parameter of the bound, with no apparent operative significance, and since the real

distortion, $\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{E}\{\rho(u_t,V_t)\}$, is then independent of $\ell$, one may maximize the lower bound w.r.t. $\ell$ over a certain set of divisors of $n$, for which $n/\ell$ is still appreciably large, such that the $o(1/\sqrt{n})$ term would remain negligible.

*Proof of Theorem 1.* First, observe that since $\hat{P}_{U^\ell X^\ell Y^\ell V^\ell Z}$ is a legitimate probability distribution, all the rules of manipulating information measures (the chain rule, condition reduces entropy, etc.) hold as usual. We will make use of the fact that $\boldsymbol{v}_i^{\ell-d}=(v_{i\ell+1},\ldots,v_{i\ell+\ell-d})$ is a deterministic function of $\boldsymbol{y}_i$ and $z_{i\ell+1}$ and therefore $(U^\ell,X^\ell)\to Y^\ell\to V^{\ell-d}$ is a Markov chain under $\hat{P}_{U^\ell X^\ell Y^\ell V^\ell|Z}$, where $V^{\ell-d}$ is random vector formed by the first $\ell-d$ components of $V^\ell$ (and similarly, below, $V_{\ell-d+1}^{\ell}$ will denote the vector formed by the remaining $d$ components). We then have the following chain of inequalities

$$\begin{aligned}
\hat{I}(U^\ell;V^{\ell-d}|Z) &\leq \hat{I}(U^\ell;Y^\ell|Z) && (12)\\
&\leq \hat{I}(U^\ell,X^\ell;Y^\ell|Z) && (13)\\
&= \hat{H}(Y^\ell|Z)-\hat{H}(Y^\ell|U^\ell,X^\ell,Z) && (14)\\
&\leq \hat{H}(Y^\ell)-\hat{H}(Y^\ell|U^\ell,X^\ell)+\hat{I}(Z;Y^\ell|U^\ell,X^\ell) && (15)\\
&\leq \hat{H}(Y^\ell)-\hat{H}(Y^\ell|U^\ell,X^\ell)+\log s. && (16)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\hat{I}(U^\ell;V^{\ell-d}|Z) &= \hat{H}(U^\ell|Z)-\hat{H}(U^\ell|V^{\ell-d},Z) && (17)\\
&\geq \hat{H}(U^\ell)-\hat{I}(Z;U^\ell)-\hat{H}(U^\ell|V^{\ell-d}) && (18)\\
&\geq \hat{H}(U^\ell)-\log s-\hat{H}(U^\ell|V^\ell)-\hat{I}(V_{\ell-d+1}^{\ell};U^\ell|V^{\ell-d}) && (19)\\
&\geq \hat{I}(U^\ell;V^\ell)-\log s-d\log M, && (20)
\end{aligned}$$

and so

$$\hat{I}(U^\ell;V^\ell)\leq\hat{H}(Y^\ell)-\hat{H}(Y^\ell|U^\ell,X^\ell)+2\log s+d\log M. \quad (21)$$

Taking now the expectation of both sides, we get

$$\begin{aligned}
\boldsymbol{E}\hat{I}(U^\ell;V^\ell) &\leq \boldsymbol{E}\hat{H}(Y^\ell)-\boldsymbol{E}\hat{H}(Y^\ell|U^\ell,X^\ell)+2\log s+d\log M\\
&\leq H(Y^\ell)-\boldsymbol{E}\hat{H}(Y^\ell|U^\ell,X^\ell)+2\log s+d\log M && (22)
\end{aligned}$$

where in the second line, $H(Y^\ell)$ is the entropy of $Y^\ell$ that is induced by $\hat{P}_{X^\ell}$ and the real channel $P_{Y^\ell|X^\ell}$. Here we have used the fact that $\hat{H}(Y^\ell)$ is a concave functional of $\hat{P}_{Y^\ell|X^\ell}$. As for the evaluation of $\boldsymbol{E}\hat{H}(Y^\ell|U^\ell, X^\ell)$, we invoke the following result (see [1], [2] and [19, Proposition 5.2] therein, as well as [6, Appendix A]): Let $\hat{P}_n$ be the first order empirical distribution associated with an $n$–sequence drawn from a memoryless $m$–ary source $P$. Then,

$$n \cdot \boldsymbol{E}D(\hat{P}_n \| P) = \frac{(m-1)\log e}{2} + o(1), \tag{23}$$

which is equivalent to

$$\boldsymbol{E}\hat{H} = H - \frac{(m-1)\log e}{2n} - o\left(\frac{1}{n}\right), \tag{24}$$

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

$$\begin{aligned}
& \boldsymbol{E}\hat{H}(Y^\ell|U^\ell, X^\ell) && (25)\\
= \;& \boldsymbol{E}\left\{\sum_{u^\ell, x^\ell} \hat{P}_{U^\ell X^\ell}(u^\ell, x^\ell)\hat{H}(Y^\ell|U^\ell = u^\ell, X^\ell = x^\ell)\right\} && (26)\\
\geq \;& \sum_{\{u^\ell, x^\ell:\ \hat{P}_{U^\ell X^\ell}(u^\ell, x^\ell) \geq \epsilon\}} \hat{P}_{U^\ell X^\ell}(u^\ell, x^\ell)\boldsymbol{E}\hat{H}(Y^\ell|U^\ell = u^\ell, X^\ell = x^\ell) && (27)\\
= \;& \sum_{\{u^\ell, x^\ell:\ \hat{P}_{U^\ell X^\ell}(u^\ell, x^\ell) \geq \epsilon\}} \hat{P}_{U^\ell X^\ell}(u^\ell, x^\ell)\left[H(Y^\ell|X^\ell = x^\ell) - \frac{(L^\ell - 1)\log e}{2n\hat{P}_{U^\ell X^\ell}(u^\ell, x^\ell)/\ell} - o\left(\frac{\ell}{n\epsilon}\right)\right] && (28)\\
\geq \;& \sum_{\{u^\ell, x^\ell:\ \hat{P}_{U^\ell X^\ell}(u^\ell, x^\ell) \geq \epsilon\}} \hat{P}_{U^\ell X^\ell}(u^\ell, x^\ell)H(Y^\ell|X^\ell = x^\ell) - \frac{\ell(JKL)^\ell \log e}{2n} - o\left(\frac{\ell}{n\epsilon}\right) && (29)\\
\geq \;& \sum_{u^\ell, x^\ell} \hat{P}_{U^\ell X^\ell}(u^\ell, x^\ell)H(Y^\ell|X^\ell = x^\ell) - \sum_{\{u^\ell, x^\ell:\ \hat{P}_{U^\ell, X^\ell}(u^\ell, x^\ell) < \epsilon\}} \hat{P}_{U^\ell X^\ell}(u^\ell, x^\ell)H(Y^\ell|X^\ell = x^\ell)\\
& - \frac{\ell(JKL)^\ell \log e}{2n} - o\left(\frac{\ell}{n\epsilon}\right) && (30)\\
\geq \;& H(Y^\ell|X^\ell) - \epsilon(JK)^\ell \cdot \ell \log L - \frac{\ell(JKL)^\ell \log e}{2n} - o\left(\frac{\ell}{n\epsilon}\right) && (31)\\
= \;& H(Y^\ell|X^\ell) - \ell \cdot \delta_0(\epsilon, \ell, n), && (32)
\end{aligned}$$

where we have defined

$$\delta_0(\epsilon, \ell, n) = \epsilon(JK)^\ell \log L + \frac{(JKL)^\ell \log e}{2n} + o\left(\frac{1}{n\epsilon}\right). \tag{33}$$

Taking $\epsilon = 1/\sqrt{n}$, we define:

$$\delta_1(\ell, n) = \delta_0\left(\frac{1}{\sqrt{n}}, \ell, n\right) = \frac{(JK)^\ell \log L}{\sqrt{n}} + \frac{(JKL)^\ell \log e}{2n} + o\left(\frac{1}{\sqrt{n}}\right). \tag{34}$$

On substituting the inequality

$$\boldsymbol{E}\hat{H}(Y^\ell|U^\ell, X^\ell) \geq H(Y^\ell|X^\ell) - \ell\delta_1(\ell, n) \tag{35}$$

into eq. (22), we get

$$\begin{aligned}
\boldsymbol{E}\hat{I}(U^\ell; V^\ell) &\leq I(X^\ell; Y^\ell) + 2\log s + d\log M + \ell\delta_1(\ell, n) && (36)\\
&\leq \ell C + 2\log s + d\log M + \ell\delta_1(\ell, n). && (37)
\end{aligned}$$

Now, denoting by $\hat{\boldsymbol{E}}$ the empirical expectation (w.r.t. $\hat{P}_{U^\ell X^\ell Y^\ell V^\ell Z}$), we obviously have

$$\begin{aligned}
\boldsymbol{E}\hat{I}(U^\ell; V^\ell) &\geq \ell \cdot \boldsymbol{E}\hat{R}_{U^\ell}\left(\frac{1}{\ell}\hat{\boldsymbol{E}}\rho(U^\ell, V^\ell)\right) && (38)\\
&= \ell \cdot \boldsymbol{E}\hat{R}_{U^\ell}\left(\frac{1}{n}\sum_{t=1}^{n}\rho(u_t, V_t)\right) && (39)\\
&\geq \ell \cdot \hat{R}_{U^\ell}\left(\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{E}\rho(u_t, V_t)\right), && (40)
\end{aligned}$$

where in the last line, we have used the convexity of the rate–distortion function. Finally, we get

$$\hat{R}_{U^\ell}\left(\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{E}\rho(u_t, V_t)\right) \leq C + \frac{2\log s + d\log M}{\ell} + \delta_1(\ell, n), \tag{41}$$

or

$$\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{E}\rho(u_t, V_t) \geq \hat{D}_{U^\ell}\left(C + \frac{2\log s + d\log M}{\ell} + \delta_1(\ell, n)\right). \tag{42}$$

This completes the proof of Theorem 1. □

While the lower bound of Theorem 1 is not quite explicit (primarily because of the complicated dependence of the function $\hat{D}_{U^\ell}(\cdot)$ on $\ell$ when $u^n$ is arbitrary), we next propose an alternative lower bound, which is simpler and more explicit. The price of this simplicity, however, is a possible loss of tightness, The idea is based on the Shannon lower bound. Suppose that $\mathcal{U} = \mathcal{V}$ is a group and the distortion measure $\rho(u, v)$ depends only on the difference $u - v$ for a well defined subtraction operation on the group (e.g., subtraction modulo $J$). Accordingly, we denote $\rho(u, v) = \varrho(v - u)$.

We define the function $\Phi(D)$ to be the maximum entropy of a random variable $W$ over an alphabet of size $J$, subject to the constraint $\boldsymbol{E}\varrho(W) \leq D$. We also define

$$\Psi(x) = \begin{cases} 0 & x < 0 \\ \Phi^{-1}(x) & x \geq 0 \end{cases} \tag{43}$$

Then, our next result is the following.

**Theorem 2** *Consider the communication setting described in Section 2. Let $u^n$ be an individual sequence, let $C$ be the capacity of the discrete memoryless channel, and let the overall coding–decoding delay be $d$. Then, for every decoder with $s$ states and a modulo–$\ell$ counter,*

$$\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{E}\{\varrho(V_t - u_t)\} \geq \Psi\left(\frac{c(u^n)\log c(u^n)}{n} - C - \frac{2\log s + d\log M}{\ell} - \delta_2(\ell, n)\right), \tag{44}$$

*where $c(u^n)$ is the number of phrases associated with incremental parsing [13] of $u^n$ and*

$$\delta_2(\ell, n) = \delta_1(\ell, n) + \frac{2\ell(1+\log J)^2}{(1-\epsilon_n)\log n} + \frac{2\ell J^{2\ell}\log J}{n} + \frac{1}{\ell}, \tag{45}$$

*$\epsilon_n$ being a positive sequence tending to zero as $n \to \infty$.*

An important feature of this bound is that the dependence on $\ell$ is now fairly explicit as it appears only in the expression $\delta_2(\ell, n) + (2\log s + d\log M)/\ell$, and so, the effect of the choice of $\ell$ can be better understood. Indeed, for decoders that are not equipped with a counter, the maximization of the bound over $\ell$, which is equivalent to the minimization of $\delta_2(\ell, n) + (2\log s + d\log M)/\ell$, is easier now. In particular, it is clear that $\ell$ should be $o(\log n)$ for this expression to vanish as $n \to \infty$. Another interesting point here is that the bound depends on $u^n$ only via its Lempel-Ziv complexity, $c(u^n)\log c(u^n)/n$. This is not a trivial fact, because the Lempel–Ziv complexity refers to the compressibility of $u^n$ using finite–state encoders, whereas here, the encoder is not limited to be a finite–state machine – only the decoder has such a limitation.

*Proof of Theorem 2.* Defining $V^\ell - U^\ell$ as the component-wise difference between the two vectors, we have:

$$\ell \cdot \hat{R}_{U^\ell}(D) = \hat{H}(U^\ell) - \max\{H(U^\ell|V^\ell):\ \boldsymbol{E}\varrho(V^\ell - U^\ell) \leq \ell D\} \tag{46}$$

$$= \hat{H}(U^\ell) - \max\{H(V^\ell - U^\ell|V^\ell):\ \boldsymbol{E}\varrho(V^\ell - U^\ell) \leq \ell D\} \tag{47}$$

$$
\begin{aligned}
&= \hat{H}(U^\ell) - \max\{H(W^\ell|V^\ell):\ \boldsymbol{E}\varrho(W^\ell) \le \ell D\} && (48)\\
&\ge \hat{H}(U^\ell) - \max\{H(W^\ell):\ \boldsymbol{E}\varrho(W^\ell) \le \ell D\} && (49)\\
&\ge \hat{H}(U^\ell) - \max\left\{\sum_{i=1}^{\ell} H(W_i):\ \sum_{i=1}^{\ell} \boldsymbol{E}\varrho(W_i) \le \ell D\right\} && (50)\\
&\ge \hat{H}(U^\ell) - \max\left\{\sum_{i=1}^{\ell} \Phi(\boldsymbol{E}\varrho(W_i)):\ \sum_{i=1}^{\ell} \boldsymbol{E}\varrho(W_i) \le \ell D\right\} && (51)\\
&\ge \hat{H}(U^\ell) - \max\left\{\ell\cdot\Phi\left(\frac{1}{\ell}\sum_{i=1}^{\ell} \boldsymbol{E}\varrho(W_i)\right):\ \sum_{i=1}^{\ell} \boldsymbol{E}\varrho(W_i) \le \ell D\right\} && (52)\\
&= \hat{H}(U^\ell) - \ell\cdot\Phi(D), && (53)
\end{aligned}
$$

where in the last two lines, we have used concavity and the monotonicity of $\Phi(\cdot)$, respectively. Now, it is shown in [5, eq. (21)] (see also [4]) that

$$
\hat{H}(U^\ell) \ge \ell\cdot\left[\frac{c(u^n)\log c(u^n)}{n} - \delta(\ell,n)\right], \tag{54}
$$

where

$$
\delta(\ell,n) = \frac{2\ell(1+\log J)^2}{(1-\epsilon_n)\log n} + \frac{2\ell J^{2\ell}\log J}{n} + \frac{1}{\ell}, \tag{55}
$$

$\epsilon_n$ being a positive sequence tending to zero, and $c(u^n)$ is the number of phrases in $u^n$ resulting from Lempel–Ziv incremental parsing. Thus,

$$
\begin{aligned}
\boldsymbol{E}\hat{R}_{U^\ell}\left(\frac{1}{n}\sum -t = 1^n\rho(V_t-u_t)\right) &\ge \frac{c(u^n)\log c(u^n)}{n} - \boldsymbol{E}\Phi\left(\frac{1}{n}\sum_{t=1}^{n}\varrho(V_t-u_t)\right) - \delta(\ell,n) && (56)\\
&\ge \frac{c(u^n)\log c(u^n)}{n} - \Phi\left(\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{E}\varrho(V_t-u_t)\right) - \delta(\ell,n). && (57)
\end{aligned}
$$

and we end up with

$$
\Phi\left(\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{E}\varrho(V_t-u_t)\right) \ge \frac{c(u^n)\log c(u^n)}{n} - C - \frac{2\log s + d\log M}{\ell} - \delta_2(\ell,n) \tag{58}
$$

or

$$
\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{E}\varrho(V_t-u_t) \ge \Psi\left(\frac{c(u^n)\log c(u^n)}{n} - C - \frac{2\log s + d\log M}{\ell} - \delta_2(\ell,n)\right). \tag{59}
$$

This completes the proof of Theorem 2. □

# 4 Distortion Bounds Under Common Reconstruction

Consider next the case where, in addition to the above–mentioned limitations on the decoder, an additional constraint is imposed, which is the constraint of almost deterministic reconstruction at

the level of $\ell$–blocks. This setting is formalized as follows. For a given vanishing sequence $\epsilon_n \in [0,1]$, we insist that

$$\boldsymbol{E}\hat{\Pr}\{V^\ell \neq \hat{V}^\ell\} \equiv \frac{\ell}{n}\sum_{i=0}^{n/\ell-1}\Pr\{V_{i\ell+1}^{i\ell+\ell} \neq \hat{v}_{i\ell+1}^{i\ell+\ell}\} \leq \epsilon_n, \tag{60}$$

where $\hat{V}^\ell = q(U^\ell)$ (and $\hat{v}_{i\ell+1}^{i\ell+\ell} = q(u_{i\ell+1}^{i\ell+\ell})$), for some deterministic function $q$, is the target reconstruction. We will assume, in this section, that $\rho_{\max} \stackrel{\Delta}{=} \max_{u,v}\rho(u,v) < \infty$. Our lower bound for this case is given by the following theorem.

**Theorem 3** *Consider the communication setting described in Section 2. Let $u^n$ be an individual sequence, let $C$ be the capacity of the discrete memoryless channel, and let the overall coding–decoding delay be $d$. Then, for every decoder with $s$ states, a modulo–$\ell$ counter and a common reconstruction constraint defined as in eq. (60):*

$$\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{E}\{\rho(u_t,V_t)\} \geq \tilde{D}_{U^\ell}\left(C + \frac{2\log s + d\log M}{\ell} + \delta_2(\ell,n) + 2\Delta(\epsilon_n)\right) - \rho_{\max}\epsilon_n, \tag{61}$$

*where $\Delta(\epsilon_n) = h_2(\epsilon_n) + \epsilon_n\ell\log J$, $h_2(\cdot)$ being the binary entropy function, and*

$$\tilde{D}_{U^\ell}(R) = \min_q\left\{\frac{1}{\ell}\hat{\boldsymbol{E}}\rho(U^\ell, q(U^\ell)) :\ \hat{H}(q(U^\ell)) \leq \ell R\right\}. \tag{62}$$

*Proof of Theorem 3.* First, under the assumption of common reconstruction (60), one readily finds, using Fano's inequality, that

$$\boldsymbol{E}\hat{H}(V^\ell|U^\ell) \leq \Delta(\epsilon_n), \tag{63}$$

where the concavity of the function $\Delta(\cdot)$ was used in order to insert the expectation into the argument of this function in order to get the real probability of error. Thus,

$$\begin{aligned}\boldsymbol{E}\hat{I}(U^\ell;V^\ell) &= \boldsymbol{E}\hat{H}(V^\ell) - \boldsymbol{E}\hat{H}(V^\ell|U^\ell) &(64)\\ &\geq \boldsymbol{E}\hat{H}(V^\ell) - \Delta(\epsilon_n). &(65)\end{aligned}$$

Now,

$$\begin{aligned}\boldsymbol{E}\hat{H}(V^\ell) &\geq \hat{H}(\hat{V}^\ell) - \boldsymbol{E}\hat{H}(\hat{V}^\ell|V^\ell) &(66)\\ &\geq \hat{H}(\hat{V}^\ell) - \Delta(\epsilon_n) &(67)\end{aligned}$$

and so,

$$\boldsymbol{E}\hat{I}(U^\ell;V^\ell) \geq \hat{H}(\hat{V}^\ell) - 2\Delta(\epsilon_n). \tag{68}$$

Now, observe that

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n}\rho(u_t,\hat{v}_t) &= \hat{\boldsymbol{E}}\left\{\frac{1}{\ell}\rho(U^\ell,q(U^\ell))\right\} && (69)\\
&= \frac{1}{\ell}\sum_{\{(u^\ell,v^\ell):\ q(u^\ell)=v^\ell\}}\hat{P}_{U^\ell,V^\ell}(u^\ell,v^\ell)\rho(u^\ell,v^\ell)+\\
&\quad\ \frac{1}{\ell}\sum_{\{(u^\ell,v^\ell):\ q(u^\ell)\neq v^\ell\}}\hat{P}_{U^\ell,V^\ell}(u^\ell,v^\ell)\rho(u^\ell,q(u^\ell)) && (70)\\
&\le \frac{1}{\ell}\sum_{u^\ell,v^\ell}\hat{P}_{U^\ell,V^\ell}(u^\ell,v^\ell)\rho(u^\ell,v^\ell)+\rho_{\max}\cdot\sum_{\{(u^\ell,v^\ell):\ q(u^\ell)\neq v^\ell\}}\hat{P}_{U^\ell,V^\ell}(u^\ell,v^\ell) && (71)\\
&= \frac{1}{n}\sum_{t=1}^{n}\rho(u_t,V_t)+\rho_{\max}\cdot\hat{\Pr}\{V^\ell\neq\hat{V}^\ell\} && (72)
\end{aligned}$$

and so, taking the expectation of both sides, we get

$$\frac{1}{n}\sum_{t=1}^{n}\rho(u_t,\hat{v}_t)\le\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{E}\rho(u_t,V_t)+\rho_{\max}\epsilon_n. \tag{73}$$

Thus, defining

$$\tilde{R}_{U^\ell}(D)=\min_q\{\frac{1}{\ell}\hat{H}(q(U^\ell)):\ \hat{\boldsymbol{E}}\rho(U^\ell,q(U^\ell))\le\ell D\}, \tag{74}$$

we readily have

$$\begin{aligned}
\boldsymbol{E}\hat{I}(U^\ell;V^\ell) &\ge \hat{H}(q(U^\ell))-2\Delta(\epsilon_n) && (75)\\
&\ge \ell\tilde{R}_{U^\ell}\left(\frac{1}{n}\sum_{t=1}^{n}\rho(u_t,\hat{v}_t)\right)-2\Delta(\epsilon_n) && (76)\\
&\ge \ell\tilde{R}_{U^\ell}\left(\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{E}\rho(u_t,V_t)+\rho_{\max}\epsilon_n\right)-2\Delta(\epsilon_n). && (77)
\end{aligned}$$

This means, of course, that

$$\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{E}\rho(u_t,V_t)\ge\tilde{D}_{U^\ell}\left(C+\frac{2\log s+d\log M}{\ell}+\delta_2(\ell,n)+\frac{2\Delta(\epsilon_n)}{\ell}\right)-\rho_{\max}\epsilon_n, \tag{78}$$

completing the proof of Theorem 3. □

Here too, performance can be expressed in terms of Lempel–Ziv complexity, as $\hat{H}(q(U^\ell))/\ell\ge[c(\hat{v}^n)\log c(\hat{v}^n)]/n-\delta'(\ell,n)$, where $\delta'(\ell,n)$ is defined just like $\delta(\ell,n)$, but with $J$ replaced by $M$. Thus,

$$\boldsymbol{E}\hat{I}(U^\ell;V^\ell)\ge\ell R_{LZ}\left(\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{E}\rho(u_t,V_t)+\rho_{\max}\epsilon_n\Big|u^n\right)-2\Delta(\epsilon_n)-\delta'(\ell,n), \tag{79}$$

where

$$R_{LZ}(D|u^n) = \min_q \left\{ \frac{c(\hat{v}^n)\log c(\hat{v}^n)}{n} : \hat{v}_{i\ell+1}^{i\ell+\ell} = q(u_{i\ell+1}^{i\ell+\ell}), \right.$$
$$\left. i = 0, 1, \ldots, n/\ell - 1, \ \frac{1}{n}\sum_{i=1}^{n} \rho(u_t, \hat{v}_t) \le D \right\}. \tag{80}$$

Note that in Section 3, we were able to get bounds on the expected distortion, thanks to the convexity of $\hat{R}_{U^\ell}(\cdot)$ and the concavity of $\Phi(\cdot)$, whereas now, we obtained such a bound by using the proximity between the actual expected distortion and the distortion between $u^n$ and its intended reconstruction $\hat{v}^n$.

# 5 Linear Encoders and Decoders

So far, we have dealt with finite alphabets only. It is possible to derive analogous results for continuous alphabets, if the encoder and decoder are limited to be linear. In this section, we provide a brief outline how this can be done, by presenting a parallel result to the Theorem 1.

Consider the following structure: The encoder is given by

$$x_t = \sum_{i=1}^{\infty} a_i x_{t-i} + \sum_{i=0}^{\infty} b_i u_{t-i}, \tag{81}$$

where $\{a_i\}$ and $\{b_i\}$ are real–valued parameters, chosen such that the encoder would satisfy a certain input constraint. The finite–state decoder we had before[4] is replaced by a decoder with the same structure, except that now $f$ and $g$ are linear functions (i.e., state–space representation):

$$v_{t-d} = \alpha z_t + \beta y_t \tag{82}$$

$$z_{t+1} = \gamma z_t + \delta y_t. \tag{83}$$

We will assume, for the sake of simplicity, that $u_t$, $x_t$, $y_t$, $v_t$ and $z_t$ are all real–valued variables (scalars), although our discussion can be generalized to the vector case ($u_t, v_t \in \mathbb{R}^k$, $x_t, y_t \in \mathbb{R}^m$, $z_t \in \mathbb{R}^p$, $k$, $m$ and $p$ positive integers), in which case, $\{a_i\}$, $\{b_i\}$, $\alpha$, $\beta$, $\gamma$ and $\delta$ become matrices of the corresponding dimensions. The channel is assumed to be a discrete–time AWGN, i.e., $Y_t = x_t + N_t$, where $N_t$ is a stationary, i.i.d. zero–mean Gaussian process with variance $\sigma^2$.

[4]For simplicity, we now refer to the one without the modulo-$\ell$ counter.

Consider first[5] the case where $\{u_t\}$ is a zero–mean, stationary Gaussian process, independent of $\{N_t\}$, and so, its notation is temporarily changed to $\{U_t\}$. Consequently, all other signals in the system become random processes, and accordingly, their notation here will use capital letters. Due to the linearity of the systems, $\{(U_t, X_t, Y_t, V_t, Z_t), \ -\infty < t < \infty\}$ are jointly Gaussian processes. We assume that these processes are jointly stationary. We also assume that the system is non–degenerated[6] in the sense that

$$\epsilon_Z^2 \stackrel{\Delta}{=} \lim_{n\to\infty} \mathrm{mmse}\{Z_1|U_1^n, X_1^n, Y_1^n\} > 0 \tag{84}$$

and similarly

$$\epsilon_V^2 \stackrel{\Delta}{=} \lim_{n\to\infty} \mathrm{mmse}\{V_0|V_{-n}^{-1}, U_{-n}^d\} > 0, \tag{85}$$

where $\mathrm{mmse}\{A|B\} = \boldsymbol{E}[A - \boldsymbol{E}(A|B)]^2$ designates the minimum mean squared error in estimating a random variable $A$ from another random variable $B$, and where the limits obviously exist due to the non–increasing monotonicity of $\mathrm{mmse}\{Z_1|U_1^n, X_1^n, Y_1^n\}$ and $\mathrm{mmse}\{V_0|V_{-n}^{-1}, U_{-n}^d\}$ as functions on $n$. The parameters $\epsilon_Z^2$ and $\epsilon_V^2$ are constants that depend on the auto-correlation function of the source, on the noise variance of noise, $\sigma^2$, and on the parameters of the encoder and decoder, $\{a_i\}$, $\{b_i\}$, $\alpha$, $\beta$, $\gamma$ and $\delta$. Obviously, $\epsilon_Z^2 \le \sigma_Z^2$ and $\epsilon_V^2 \le \sigma_V^2$, where $\sigma_Z^2$ and $\sigma_V^2$ are the variances of $Z_t$ and $V_t$, respectively. We define $U^\ell = (U_1, \ldots, U_\ell)$, $X^\ell = (X_1, \ldots, X_\ell)$, $Y^\ell = (Y_1, \ldots, Y_\ell)$, $V^\ell = (V_1, \ldots, V_\ell)$, and $Z = Z_1$. We begin similarly as in eqs. (12), but the last step must be modified slightly:

$$\begin{aligned}
I(U^\ell; V^{\ell-d}|Z) &\le I(U^\ell; Y^\ell|Z) && (86)\\
&\le I(U^\ell, X^\ell; Y^\ell|Z) && (87)\\
&= h(Y^\ell|Z) - h(Y^\ell|U^\ell, X^\ell, Z) && (88)\\
&\le h(Y^\ell) - h(Y^\ell|U^\ell, X^\ell) + I(Z; Y^\ell|U^\ell, X^\ell) && (89)\\
&= h(Y^\ell) - h(Y^\ell|X^\ell) + \frac{1}{2}\log\frac{\mathrm{mmse}\{Z|U^\ell, X^\ell\}}{\mathrm{mmse}\{Z|U^\ell, X^\ell, Y^\ell\}} && (90)\\
&\le I(X^\ell; Y^\ell) + \frac{1}{2}\log\frac{\sigma_Z^2}{\epsilon_Z^2} && (91)\\
&\le \ell C + \frac{1}{2}\log\frac{\sigma_Z^2}{\epsilon_Z^2}. && (92)
\end{aligned}$$

[5]This assumption will be dropped soon.

[6]For example, if $\gamma = \delta = 0$ and hence $Z_t \equiv 0$, or if $\alpha = \beta = 0$ and hence $V_t \equiv 0$, the system is obviously degenerated.

On the other hand,

$$\begin{aligned} I(U^\ell;V^{\ell-d}|Z) &= h(U^\ell|Z)-h(U^\ell|V^{\ell-d},Z) && (93)\\ &\geq h(U^\ell|Z)-h(U^\ell|V^{\ell-d}) && (94)\\ &= h(U^\ell)-I(Z;U^\ell)-h(U^\ell|V^\ell)-I(V_{\ell-d+1}^\ell;U^\ell|V^{\ell-d}) && (95)\\ &= h(U^\ell)-h(U^\ell|V^\ell)-I(Z;U^\ell)-\sum_{i=\ell-d+1}^{\ell} I(V_i;U^\ell|V^{i-1}) && (96)\\ &\geq I(U^\ell;V^\ell)-\frac{1}{2}\log\frac{\sigma_Z^2}{\epsilon_Z^2}-\frac{d}{2}\log\frac{\sigma_V^2}{\epsilon_V^2}, && (97) \end{aligned}$$

and so,

$$I(U^\ell;V^\ell)\leq \ell C+\log\frac{\sigma_Z^2}{\epsilon_Z^2}+\frac{d}{2}\log\frac{\sigma_V^2}{\epsilon_V^2}. \tag{98}$$

This is quite analogous to the bounds we obtained in the finite–alphabet case, but now $\log s$ and $\log M$ are replaced by $\log\frac{\sigma_Z}{\epsilon_Z}$ and $\log\frac{\sigma_V}{\epsilon_V}$, respectively, thus $\frac{\sigma_Z}{\epsilon_Z}$ and $\frac{\sigma_V}{\epsilon_V}$ play roles of effective alphabet sizes (or effective resolution levels) of the variables $Z_t$ and $V_t$, respectively. Now, clearly, in the Gaussian case, $I(U^\ell;V^\ell)$ depends on the joint density of $(U^\ell,V^\ell)$ only via the covariance matrix of this random vector. Equivalently, consider the class of Gaussian channels from $U^\ell$ to $V^\ell$, defined by

$$V^\ell=GU^\ell+W^\ell \tag{99}$$

where $G$ is a deterministic $\ell\times\ell$ matrix and $W^\ell$ is a zero–mean Gaussian vector, independent of $U^\ell$, with covariance matrix $\Sigma_W$. Denoting the covariance matrix of $U^\ell$ by $\Sigma_U$, then

$$I(U^\ell;V^\ell)=\frac{1}{2}\log\frac{\det(G\Sigma_UG^T+\Sigma_W)}{\det(\Sigma_W)}=\frac{1}{2}\log\det(I+\Sigma_W^{-1}G\Sigma_UG^T) \tag{100}$$

Thus, defining

$$R_\ell(D)=\min_{G,\Sigma_W}\left\{\frac{1}{2\ell}\log\det(I+\Sigma_W^{-1}G\Sigma_UG^T):\ \boldsymbol{E}\rho(U^\ell,V^\ell)\leq\ell D\right\}, \tag{101}$$

where $\rho$ designates the quadratic distortion measure (or any other distortion measure that such that $\boldsymbol{E}\rho(U^\ell,V^\ell)$ depends only on the covariance matrix of $(U^\ell,V^\ell)$), we have

$$R_\ell(D)\leq C+\frac{1}{\ell}\log\frac{\sigma_Z^2}{\epsilon_Z^2}+\frac{d}{2\ell}\log\frac{\sigma_V^2}{\epsilon_V^2}, \tag{102}$$

or

$$\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{E}\rho(U_t,V_t)\geq D_\ell\left(C+\frac{1}{\ell}\log\frac{\sigma_Z^2}{\epsilon_Z^2}+\frac{d}{2\ell}\log\frac{\sigma_V^2}{\epsilon_V^2}\right), \tag{103}$$

Now, the l.h.s. of (102) depends only on the covariance matrix of the source, whereas $C$ (or $I(X^\ell;Y^\ell)$) depends only on the covariance matrix $\Sigma_X$ of $X^\ell$ and the covariance matrix $\Sigma_N$ of the noise vector, which we have taken to be $\sigma^2 I$. Since the encoding and decoding systems are linear, the auto-correlation cross–correlation functions of their outputs depend only on those of their inputs (for a given linear encoder and decoder), no matter whether these processes are Gaussian or not. The expected distortion also depends on the joint density of $U^\ell, V^\ell$) only via the variances and covariances of their components. Consequently, at this point, the Gaussian assumption becomes immaterial. The source $U^\ell$ may have any pdf with a given covariance matrix $\Sigma_U$. In particular, we can take $\Sigma_U$ to be the empirical covariance matrix of a deterministic source sequence $u^n$. In this case, in the above chains of inequalities, all information measures should be replaced by their empirical counterparts, which depend on the empirical covariances instead of the true covariances. The only exception is that, similarly as in the finite alphabet case, in eq. (86), it is no longer true that $\hat{h}(Y^\ell|X^\ell,U^\ell)=\hat{h}(Y^\ell|X^\ell)$, since there might be empirical correlations between the source vector and the noise vector. However, $\boldsymbol{E}\hat{h}(Y^\ell|X^\ell,U^\ell)$ tends to $h(Y^\ell|X^\ell)$ by the weak law of large numbers, so as before, upon taking expectations, one can obtain a distortion bound analogous to the one we obtained in the finite–alphabet case. In particular, for the quadratic distortion measure, we have:

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{E}\rho(u_t,V_t) &\geq \min_{G,\Sigma_W}\left\{\frac{1}{n}\sum_{i=0}^{n/\ell-1}\boldsymbol{E}\rho(u_{i\ell+1}^{i\ell+\ell},Gu_{i\ell+1}^{i\ell+\ell}+W_{i\ell+1}^{i\ell+\ell}):\right.\\
&\qquad\left.\frac{1}{2\ell}\log\det(I+\Sigma_W^{-1}G\hat{\Sigma}_UG^T)\leq C+\frac{1}{\ell}\log\frac{\sigma_Z^2}{\epsilon_Z^2}+\frac{d}{2\ell}\log\frac{\sigma_V^2}{\epsilon_V^2}+\epsilon_n\right\} \quad (104)\\
&= \min_{G,\Sigma_W}\left\{\mathrm{tr}\{(G-I)\hat{\Sigma}_U(G^T-I)+\frac{1}{\ell}\Sigma_W\}:\right.\\
&\qquad\left.\frac{1}{2\ell}\log\det(I+\Sigma_W^{-1}G\hat{\Sigma}_UG^T)\leq C+\frac{1}{\ell}\log\frac{\sigma_Z^2}{\epsilon_Z^2}+\frac{d}{2\ell}\log\frac{\sigma_V^2}{\epsilon_V^2}+\epsilon_n\right\}, \quad (105)
\end{aligned}$$

where $\hat{\Sigma}_U=\frac{\ell}{n}\sum_{i=0}^{n/\ell-1}\boldsymbol{u}_i\boldsymbol{u}_i^T$ is the empirical covariance of the source, $W_{i\ell+1}^{i\ell+\ell}$ is a zero–mean random vector with covariance matrix $\Sigma_W$ and $\epsilon_n$ is the vanishing difference between $\boldsymbol{E}\hat{h}(Y^\ell|X^\ell,U^\ell)/\ell$ and $h(Y|X)$. The point here is that for the purpose of obtaining a lower bound on the distortion attainable by linear encoders and decoders, we are replacing the optimization over infinitely many parameters $\{a_i\}$, $\{b_i\}$, $\alpha$, $\beta$, $\gamma$, and $\delta$, by optimization over two $\ell\times\ell$ matrices, $G$ and $\Sigma_W$, at the possible rate loss of $\frac{1}{\ell}\log\frac{\sigma_Z^2}{\epsilon_Z^2}+\frac{d}{2\ell}\log\frac{\sigma_V^2}{\epsilon_V^2}+\epsilon_n$, which vanishes as $\ell$ and $n$ grow. Thus, the parameter

$\ell$ trades off the quality of the bound (its tightness) with the complexity of the optimization.

Note that here our bounds are a bit weaker than in the finite–alphabet case, in the sense that they depend on the competing linear system with parameters $\{a_i\}$, $\{b_i\}$, $\alpha$, $\beta$, $\gamma$ and $\delta$ (via $\epsilon_V^2$ and $\epsilon_Z^2$). However, the dependence on these parameters becomes weaker and weaker as $\ell$ grows without bound.

# Appendix

*Some Concerns About the Proof of Theorem 3 in [11].*

First, it should be pointed out that in [11, p. 140], the encoder was also assumed to be a finite–state machine, and so, in this appendix, following the notation of [11], the state of the encoder is denoted by $z_t$ and the state of the decoder is denoted by $z'_t$.

In [11], the joint probability distribution of all random variables was defined (in our notation) to be

$$\begin{aligned} & \hat{P}_{U^\ell X^\ell Y^\ell V^\ell Z Z'}(u^\ell, x^\ell, y^\ell, v^\ell, z, z') \\ = \;& P(z, z')\hat{P}_{U^\ell}(u^\ell)\hat{P}_{X^\ell|U^\ell,Z}(x^\ell|u^\ell, z)P(y^\ell|x^\ell)\hat{P}_{V^\ell|X^\ell,Z'}(v^\ell|x^\ell, z'), \end{aligned} \tag{A.1}$$

where $P(z, z')$ is the expectation of the joint empirical distribution of the state of the encoder, denoted here by $Z$, and the state of the decoder, denoted here by $Z'$, at the beginnings of all $\ell$-blocks, and $P(y^\ell|x^\ell)$ is the *real* conditional probability associated with the channel. First, observe that according to this definition, $U^\ell$ is taken to be independent of $Z$ and $Z'$, which is inconsistent with the fact that the encoder state $Z$ varies in response to the source and that there might be empirical dependencies between successive $\ell$–blocks of the source. Also, according to this definition, $Y^\ell$ is independent of $Z'$ given $X^\ell$, which similarly to the earlier comment, does not seem to settle with the fact that $Z'$ responds to the decoder input $Y^\ell$.

Another issue is the use of the data processing theorem when it comes to empirical distributions. For example, the equality [11, p. 141, top] $\hat{I}(Z, U^\ell, X^\ell; V^\ell) = \hat{I}(Z, X^\ell; V^\ell)$ is questionable because there might be incidental empirical dependencies between $U^\ell$ and $V^\ell$ given $(Z, X^\ell)$.

Finally, we have concerns regarding the way in which the delay was handled in [11], where the decoder output $v_{t-d}$ was simply renamed $v_t$. It should be kept in mind that while the data

processing theorem applies to $l$–blocks of $\{u_t\}$, $\{x_t\}$, $\{y_t\}$ and $\{v_{t-d}\}$, the distortion is measured between $u_t$ and $v_t$, and so, the discrepancy between the $\{v_t\}$ and its delayed version $\{v_{t-d}\}$ is real and cannot be handled by simple renaming. Indeed, in [11], the lower bound does not depend on $d$, a fact which is in contrast to the expectation that the larger is $d$, the better is the performance that can be achieved.

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

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

[5] N. Merhav, "Perfectly secure encryption of individual sequences," *IEEE Trans. Inform. Theory*, vol. 59, no. 3, pp. 1302–1310, March 2013.

[6] 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.

[7] 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.

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

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

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

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

[12] 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.

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