A Universal Random Coding Ensemble for Sample-wise Lossy Compression

Neri Merhav

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

Abstract

We propose a universal ensemble for random selection of rate-distortion codes, which is asymptotically optimal in a sample-wise sense. According to this ensemble, each reproduction vector, $\hat{\boldsymbol{x}}$, is selected independently at random under the probability distribution that is proportional to $2^{-LZ(\hat{\boldsymbol{x}})}$, where $LZ(\hat{\boldsymbol{x}})$ is the code-length of $\hat{\boldsymbol{x}}$ pertaining to the 1978 version of the Lempel-Ziv (LZ) algorithm. We show that, with high probability, the resulting codebook gives rise to an asymptotically optimal variable-rate lossy compression scheme under an arbitrary distortion measure, in the sense that a matching converse theorem also holds. According to the converse theorem, even if the decoder knew ℓ-th order type of source vector in advance (ℓ being a large but fixed positive integer), the performance of the above-mentioned code could not have been improved essentially, for the vast majority of codewords that represent all source vectors in the same type. Finally, we provide a discussion of our results, which includes, among other things, a comparison to a coding scheme that selects the reproduction vector with the shortest LZ code length among all vectors that are within the allowed distortion from the source vector.

Index Terms: lossy compression, rate-distortion theory, Lempel-Ziv algorithm, universal coding, sphere covering.

1 Introduction

We revisit the well-known problem of lossy source coding for finite-alphabet sequences with respect to (w.r.t.) a certain distortion measure [2], [4, Chap. 10], [7, Chap. 9], [9], [27, Chaps. 7,8]. More concretely, our focus is on d-semifaithful codes, namely, variable–rate codes that meet a certain distortion constraint for every source sequence (and not only in expectation). As is very well known [2], the rate-distortion function quantifies the minimum achievable expected coding rate for a given memoryless source and distortion measure.

During several past decades, many research efforts were motivated by the fact that the source statistics are rarely (if not never) known in practice, and were therefore directed to the quest for universal coding schemes, namely, coding schemes which do not depend of the unknown statistics, but nevertheless, approach the lower bounds (i.e., the entropy, in lossless compression, or the rate-distortion function, in the lossy case) asymptotically, as the block length grows without bound. We next provide a very brief (and non-comprehensive) review of some of the relevant earlier works.

In lossless compression, the theory of universal source coding is very well developed and mature. Davisson's work [6] concerning universal-coding redundancies has established the concepts of weak universality and strong universality (vanishing maximin and minimax redundancies, respectively), and has characterized the connection to the capacity of the 'channel' defined by family of conditional distributions of the data to be compressed given the index (or parameter) of the source in the class [8]. For many of the frequently encountered parametric classes of sources, the minimum achievable redundancy of universal codes is well-known to be dominated by $\frac{k\log n}{2n}$, where k is the number of degrees of freedom of the parameter, and n is the block length. A central idea that arises from Davisson's theory is to construct a Shannon code pertaining to the probability distribution of the data vector w.r.t. a mixture (with a certain prior function) of all sources in the class. Rissanen, which was the inventor of the minimum description length (MDL) principle, has proved in [25] a converse to a coding theorem, which asserts that asymptotically, no universal code can achieve redundancy below $(1-\epsilon)\frac{k\log n}{2n}$, with the possible exception of sources from a subset of the parameter space, whose volume tends to zero as $n\to\infty$, for every positive ϵ. Merhav and Feder [21] have generalized this result to more general classes of sources, with the term $\frac{k\log n}{2n}$ substituted by the capacity of the above mentioned 'channel'. Further developments, including more refined redundancy analyses, have been carried out in later studies.

In the wider realm of universal lossy compression, the theory is, unfortunately, not as sharp and well-developed as in the lossless setting. We confine our attention, in this work, to d-semifaithful codes [24], namely, codes that satisfy the distortion requirement with probability one. Zhang, Yang and Wei [29] have proved that, unlike in lossless compression, in the lossy case, even if the source statistics are known perfectly, it is impossible to achieve redundancy below $\frac{\log n}{2n}$ (see also [15]), but $\frac{\log n}{n}$ is achievable. Not knowing the source conveys the price of enlarging the multiplicative constant in front of $\frac{\log n}{n}$. Indeed, Yu and Speed [28] have established weak universality with a constant that grows with the cardinalities of the alphabets of the source and the reconstruction [26]. Ornstein and Shields [24] have considered universal d-semifaithful coding for stationary and ergodic sources w.r.t. the Hamming distortion measure, and established convergence with probability one to the rate-distortion function. Kontoyiannis [10] had several interesting findings. The first is a certain central limit theorem (CLT), with a $O(1/\sqrt{n})$ redundancy term, whose coefficient is described as a limiting Gaussian random variable with some constant variance. The second is the so called law of iterated logarithm (LIL) with redundancy proportional to $\sqrt{\frac{\log(\log n)}{n}}$ infinitely often with probability one. One of the counter-intuitive conclusions from [10] is that universality is priceless under these performance measures. In [11], many of the findings are based on the observation that optimal compression can be characterized in terms of the negative logarithm of the probability of a sphere of radius nD around the source vector w.r.t. the distortion measure, where D is the allowed per-letter distortion. In the same article, they proposed also the ensemble of random coding w.r.t. a probability distribution given by a mixture of all distributions in a certain class. In two recent articles, Mahmood and Wagner [12], [13] have studied d-semifaithful codes that are strongly universal w.r.t. both the source and the distortion function. The redundancy rates in [12] behave like $\frac{\log n}{n}$ with different multiplicative constants.

A parallel line of research work on universal lossless and lossy compression, which was pioneered by Ziv, pertains to the individual-sequence approach. According to this approach, there are no assumptions on the statistical properties of the source. The source sequence to be compressed is considered an arbitrary deterministic (individual) sequence, but limitations are imposed on the encoder and/or the decoder to be implementable by finite–state machines. This includes, first and foremost, the celebrated Lempel-Ziv (LZ) algorithm [30], [33], as well as further developments that extend the scope to lossy compression with and without side information [22], [32], as well as to joint source–channel coding [16], [18], [19], [31]. In the lossless case, the article [23] provides an individual-sequence analogue of the above-mentioned result due to Rissanen, where the expression $\frac{k \log n}{2n}$ continues to designate the best achievable redundancy, but the main term of the compression ratio there is the empirical entropy of the source vector instead of the ordinary entropy of the probabilistic setting. The converse bound of [23] applies to the vast majority of source sequences within each type, and the vast majority of types (in analogy to the vast majority of the parameter space in Rissanen's framework). In a way, this kind of a converse result still contains some flavor of the probabilistic setting, because arguing that the number of exceptional typical sequences is relatively small, is actually equivalent to imposing a uniform distribution across the type and asserting that the induced probability of violating the bound is small. A similar comment applies, of course, to the exclusion of a minority of the types. The achievability result of [23], on the other hand, holds pointwise, for every sequence. A similar comment applies to [14], where asymptotically pointwise lossy compression was established with respect to first order statistics (i.e., "memoryless" statistics) with an emphasis on distortion-universality, similarly as in [12] and [13].

A similar kind of a mix between the probabilistic setting and the individual-sequence setting is adopted in this paper as well, in the context of universal rate-distortion coding, but here, just like in [**?**], there is no limitation to finite-state encoders/decoders as in [23]. In particular, our converse theorem asserts that given an arbitrary variable-rate code, and given an arbitrary distortion function within a certain wide class, the majority of reproduction vectors that represent source sequences of a given type (of any fixed order), must have a code-length that is essentially at least as large as the negative logarithm of the probability of a ball with normalized radius D (D being the allowed per-letter distortion), centered at the given source sequence. The probability of this ball is taken w.r.t. a universal distribution that is proportional to $2^{-LZ(\hat{\boldsymbol{x}})}$, where $LZ(\hat{\boldsymbol{x}})$ is the code-length of LZ encoding of the reproduction vector, $\hat{\boldsymbol{x}}$. On the other hand, we also present a matching achievability result, asserting that for every source sequence, this code length is essentially achievable by random coding, using a universal ensemble of codes, which is defined by independent random selection, where each codeword is drawn under the above-described universal probability distribution.

While the achievability result in [14] was pointwise as well, it was tailored to a memoryless structure in the sense that it was given in terms of the rate-distortion function of the first-order empirical distribution, which is blind to any empirical dependencies and repetitive patterns within the source sequence. In this paper, we both extend the scope to general individual sequences beyond the memoryless statistics and extend the allowable class of distortion measures. In terms of the technical aspects, the proof of the achievablity result is very similar to the parallel proof in [14], but the novelty lies considerably more in the converse theorem and its proof.

The outline of this paper is as follows. In Section 2, we establish the notation conventions, define a few terms and quantities, and provide some background. In Section 3, we present the converse theorem and its proof. In Section 4, we present the achievability theorem and prove it. Finally, in Section 5, we summarize the paper and discuss our results.

2 Notation, Definitions and Background

Throughout the paper, random variables will be denoted by capital letters, specific values they may take will be denoted by the corresponding lower case letters, and their alphabets will be denoted by calligraphic letters. Random vectors and their realizations will be denoted, respectively, by capital letters and the corresponding lower case letters, both in the bold face font. Their alphabets will be superscripted by their dimensions. The source vector of length n, $(x_1, \ldots, x_n)$, with components, x_i, $i = 1, \ldots, n$, from a finite-alphabet, $\mathcal{X}$, will be denoted by $\boldsymbol{x}$. The set of all such n-vectors will be denoted by $\mathcal{X}^n$, which is the n–th order Cartesian power of $\mathcal{X}$. Likewise, a reproduction vector of length n, $(\hat{x}_1, \ldots, \hat{x}_n)$, with components, $\hat{x}_i$, $i = 1, \ldots, n$, from a finite-alphabet, $\hat{\mathcal{X}}$, will be denoted by $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}^n$. We denote the cardinalities of $\mathcal{X}$ and $\hat{\mathcal{X}}$ by J and K, respectively.

For $i \leq j$, the notation x_i^j will be used to denote the substring $(x_i, x_{i+1}, \ldots, x_j)$. Probability distributions will be denoted by the letter P or Q with possible subscripts, depending on the context. The probability of an event $\mathcal{E}$ will be denoted by $\Pr\{\mathcal{E}\}$, and the expectation operator with respect to (w.r.t.) a probability distribution P will be denoted by $\boldsymbol{E}\{\cdot\}$. For two positive sequences, a_n and b_n, the notation $a_n \doteq b_n$ will stand for equality in the exponential scale, that is, $\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0$. Similarly, $a_n \stackrel{\cdot}{\leq} b_n$ means that $\limsup_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} \leq 0$, and so on. The indicator function of an event $\mathcal{E}$ will be denoted by $\mathcal{I}\{E\}$. The notation $[x]_+$ will stand for $\max\{0, x\}$. The logarithmic function, $\log x$, will be understood to be defined to the base 2. Logarithms to the base e will be denote by ln.

Let ℓ be a positive integer that divides n. The ℓth order empirical distribution of $\boldsymbol{x} \in \mathcal{X}^n$, which will be denoted by $\hat{P}^\ell_{\boldsymbol{x}}$, is the vector of relative frequencies $\{\hat{P}^\ell_{\boldsymbol{x}}(a^\ell),\ a^\ell \in \mathcal{X}^\ell\}$, where

$$\hat{P}^\ell_{\boldsymbol{x}}(a^\ell) = \frac{\ell}{n} \sum_{i=0}^{n/\ell-1} \mathcal{I}\{x_{i\ell+1}^{(i+1)\ell} = a^\ell\}. \tag{1}$$

The set of all ℓth order empirical distributions of sequences in $\mathcal{X}^n$ will be denoted by $\mathcal{P}^\ell_n$. For $P^\ell \in \mathcal{P}^\ell_n$, the type class, $\{\boldsymbol{x} \in \mathcal{X}^n :\ \hat{P}^\ell_{\boldsymbol{x}} = P^\ell\}$, will be denoted by $\mathcal{T}_n(P^\ell)$. Likewise, $\mathcal{T}_n(Q^\ell)$ will denote $\{\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}^n :\ \hat{P}^\ell_{\hat{\boldsymbol{x}}} = Q^\ell\}$, where $\hat{P}^\ell_{\hat{\boldsymbol{x}}}$ is the ℓ-th order empirical distribution of $\hat{\boldsymbol{x}}$. Finally, $\hat{P}^\ell_{\boldsymbol{x}\hat{\boldsymbol{x}}}$

will denote the ℓth order joint empirical distribution of $(\boldsymbol{x}, \hat{\boldsymbol{x}})$, i.e.,

$$\hat{P}^{\ell}_{\boldsymbol{x}\hat{\boldsymbol{x}}}(a^{\ell}, b^{\ell}) = \frac{\ell}{n} \sum_{i=0}^{n/\ell-1} \mathcal{I}\{x_{i\ell+1}^{(i+1)\ell} = a^{\ell}, \hat{x}_{i\ell+1}^{(i+1)\ell} = b^{\ell}\}, \quad (a^{\ell}, b^{\ell}) \in \mathcal{X}^{\ell} \times \hat{\mathcal{X}}^{\ell}. \tag{2}$$

For a given positive integer n, a distortion function, d, is a function from $\mathcal{X}^n \times \hat{\mathcal{X}}^n$ into $\mathrm{I\!R}^+$. In the two main parts of this paper, different assumptions will be imposed on the distortion function.

1. For the achievability theorem, the distortion function can be completely arbitrary.

2. For the converse theorem, we assume that $d(\boldsymbol{x}, \hat{\boldsymbol{x}})$ depends on $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ only via their first order joint empirical distribution, $\hat{P}^1_{\boldsymbol{x}\hat{\boldsymbol{x}}}$, and that for a given such distribution, it grows linearly in n, that is, $d(\boldsymbol{x}, \hat{\boldsymbol{x}}) = n\rho(\hat{P}^1_{\boldsymbol{x}\hat{\boldsymbol{x}}})$, where the function ρ is independent of n.

Regarding item 2, additive distortion measures, which obviously comply with the requirement, are given by linear functionals of $\hat{P}^1_{\boldsymbol{x}\hat{\boldsymbol{x}}}$. However, here arbitrary non-linear functionals are allowed as well.

A rate-distortion block code of length n is a mapping, $\phi_n : \mathcal{X}^n \to \mathcal{B}_n$, $\mathcal{B}_n \subset \{0,1\}^*$, that maps the space of source vectors of length n, $\mathcal{X}^n$, into a set, $\mathcal{B}_n$, of variable-length compressed bit strings. The decoder is a mapping $\psi_n : \mathcal{B}^n \to \mathcal{C}_n \subseteq \hat{\mathcal{X}}^n$ that maps the set of compressed variable-length binary strings into a reproduction codebook, $\mathcal{C}^n$. A block code is called d-semifaithful if for every $\boldsymbol{x} \in \mathcal{X}^n$, $d(\boldsymbol{x}, \psi_n(\phi_n(\boldsymbol{x}))) \leq nD$. The code-length for $\boldsymbol{x}$, denoted $L(\boldsymbol{x})$, is the number of bits of $\phi_n(\boldsymbol{x})$. Since $L(\boldsymbol{x})$ depends on $\boldsymbol{x}$ only via $\phi_n(\boldsymbol{x})$, we will also denote it sometimes as $L(\phi_n(\boldsymbol{x}))$ or by $L(\hat{\boldsymbol{x}})$ ($\hat{\boldsymbol{x}}$ being the reproduction vector pertaining to $\phi_n(\boldsymbol{x})$), with a slight abuse of notation. For the converse theorem, we assume that correspondence between $\mathcal{B}_n$ and $\mathcal{C}_n$ is one-to-one. For the achievability theorem, we consider prefix-free codes. Accordingly, the encoder can equivalently be presented as a cascade of a reproduction encoder (a.k.a. vector quantizer), which maps $\mathcal{X}^n$ into $\mathcal{C}_n$, followed by an entropy coder, which maps $\mathcal{C}_n$ into $\mathcal{B}_n$ with no additional loss of information.

For the purpose of presenting both the converse theorem and the achievability theorem, we need to recall a few terms and facts concerning the 1978 version of LZ algorithm (a.k.a. the LZ78 algorithm) [33]. The incremental parsing procedure of the LZ78 algorithm is a procedure of sequentially parsing a vector, $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}^n$, such that each new phrase is the shortest string that has not been encountered before as a parsed phrase, with the possible exception of the last phrase, which might be incomplete. For example, the incremental parsing of the vector $\hat{\boldsymbol{x}}=$ abbabaabbaaabaa is a,b,ba,baa,bb,aa,ab,aa. Let $c(\hat{\boldsymbol{x}})$ denote the number of phrases in $\hat{\boldsymbol{x}}$ resulting from the incremental parsing procedure. Let $LZ(\hat{\boldsymbol{x}})$ denote the length of the LZ78 binary compressed code for $\hat{\boldsymbol{x}}$. According to [33, Theorem 2],

$$\begin{aligned}
LZ(\hat{\boldsymbol{x}}) &\le [c(\hat{\boldsymbol{x}})+1]\log\{2K[c(\hat{\boldsymbol{x}})+1]\} \\
&= c(\hat{\boldsymbol{x}})\log[c(\hat{\boldsymbol{x}})+1]+c(\hat{\boldsymbol{x}})\log(2J)+\log\{2K[c(\hat{\boldsymbol{x}})+1]\} \\
&= c(\hat{\boldsymbol{x}})\log c(\hat{\boldsymbol{x}})+c(\hat{\boldsymbol{x}})\log\left[1+\frac{1}{c(\hat{\boldsymbol{x}})}\right]+c(\hat{\boldsymbol{x}})\log(2K)+\log\{2K[c(\hat{\boldsymbol{x}})+1]\} \\
&\le c(\hat{\boldsymbol{x}})\log c(\hat{\boldsymbol{x}})+\log e+\frac{n(\log K)\log(2K)}{(1-\epsilon_n)\log n}+\log[2K(n+1)] \\
&\stackrel{\triangle}{=} c(\hat{\boldsymbol{x}})\log c(\hat{\boldsymbol{x}})+n\cdot\epsilon(n),
\end{aligned} \tag{3}$$

where we remind that K is the cardinality of $\hat{\mathcal{X}}$, and where $\epsilon(n)$ clearly tends to zero as $n\to\infty$, at the rate of $1/\log n$. We next define a *universal probability distribution* (see also [3], [20]):

$$U(\hat{\boldsymbol{x}})=\frac{2^{-LZ(\hat{\boldsymbol{x}})}}{\sum_{\hat{\boldsymbol{x}}'\in\hat{\mathcal{X}}^n}2^{-LZ(\hat{\boldsymbol{x}}')}}, \qquad \hat{\boldsymbol{x}}\in\hat{\mathcal{X}}^n. \tag{4}$$

Finally, we define the D-sphere around $\boldsymbol{x}$ as

$$\mathcal{S}(\boldsymbol{x},D)=\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}, \tag{5}$$

and

$$U[\mathcal{S}(\boldsymbol{x},D)]=\sum_{\hat{\boldsymbol{x}}\in\mathcal{S}(\boldsymbol{x},D)}U(\hat{\boldsymbol{x}}). \tag{6}$$

For later use, we also define

$$\hat{\mathcal{S}}(\hat{\boldsymbol{x}},D)=\{\boldsymbol{x}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}. \tag{7}$$

Our purpose is to derive upper and lower bounds on the smallest achievable code length, $L(\boldsymbol{x})$, for d-semifaithful block codes of length n, and individual sequences, $\{\boldsymbol{x}\}$, from a given ℓth order type class, $\mathcal{T}_n(P^\ell)$. As will be seen shortly, in both the converse and the achievability theorems, the main term of the bound on the length function will be $-\log(U[\mathcal{S}(\boldsymbol{x},D)])$.

3 The Converse Theorem

The following converse theorem asserts that even if the type class of the source vector was known to the decoder ahead of time, the code length could not be much smaller than $-\log(U[\mathcal{S}(\boldsymbol{x},D)])$

for the vast majority of the codewords pertaining to that type.

Theorem 1 *Let ℓ be a positive integer that divides n and let $\hat{P}^\ell$ be an arbitrary empirical distribution pertaining to a certain type class, $\mathcal{T}_n(\hat{P}^\ell)$, of source sequences in $\mathcal{X}^n$. Let d be a distortion function that depends on $(\boldsymbol{x},\hat{\boldsymbol{x}})$ only via $\hat{P}_{\boldsymbol{x}\hat{\boldsymbol{x}}}$. Then, for every every d-semifaithful variable-length block code, with one-to-one correspondence between $\mathcal{B}_n$ and $\mathcal{C}_n$, and for every $\epsilon>0$, the following lower bound applies to a fraction of at least $(1-2n^{-\epsilon})$ of the codewords, $\{\phi_n(\boldsymbol{x}),~\boldsymbol{x}\in\mathcal{T}_n(\hat{P}^\ell)\}$:*

$$L(\phi_n(\boldsymbol{x}))\ge -\log(U[\mathcal{S}(\boldsymbol{x},D)])-n\Delta_n(\ell)-\epsilon\log n, \tag{8}$$

where $\Delta_n(\ell)$ has the property $\lim_{n\to\infty}\Delta_n(\ell)=1/\ell$.

As a technical note, observe that $\Delta_n(\ell)$ can be made small only when ℓ is chosen large, as $\Delta_n(\ell)$ behaves like $1/\ell$ for fixed ℓ and large n. This suggests that the theorem is meaningful mainly when ℓ is appreciably large, which is not surprising, because the larger is ℓ, the better one can exploit empirical dependencies within the source sequence.

The remaining part of this section is devoted to the proof of Theorem 1.

Proof. We first establish a relationship that will be used later on. For two given types $\mathcal{T}_n(P^\ell)\subset\mathcal{X}^n$ and $\mathcal{T}_n(Q^\ell)\subset\hat{\mathcal{X}}^n$, consider the quantity,

$$N(D)=\sum_{\boldsymbol{x},\hat{\boldsymbol{x}}}\mathcal{I}\{\boldsymbol{x}\in\mathcal{T}_n(P^\ell),~\hat{\boldsymbol{x}}\in\mathcal{T}_n(Q^\ell),~d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}. \tag{9}$$

We can evaluate $N(D)$ in two ways. The first is as follows:

$$\begin{aligned} N(D) &= \sum_{\boldsymbol{x}\in\mathcal{T}_n(P^\ell)}\left|T_n(Q^\ell)\bigcap\mathcal{S}(\boldsymbol{x},D)\right| && (10)\\ &= |\mathcal{T}_n(P^\ell)|\cdot\left|T_n(Q^\ell)\bigcap\mathcal{S}(\boldsymbol{x},D)\right|, && (11)\end{aligned}$$

where the second equality is since $|T_n(Q^\ell)\bigcap\mathcal{S}(\boldsymbol{x},D)\Big|$ is the same for all $\boldsymbol{x}\in\mathcal{T}_n(P^\ell)$, due to the permutation-invariance assumption on the distortion function. By the same token, we can also express $N(D)$ in the following manner:

$$\begin{aligned} N(D) &= \sum_{\hat{\boldsymbol{x}}\in\mathcal{T}_n(Q^\ell)}\left|T_n(P^\ell)\bigcap\hat{\mathcal{S}}(\hat{\boldsymbol{x}},D)\right| && (12)\\ &= |\mathcal{T}_n(Q^\ell)|\cdot\left|T_n(P^\ell)\bigcap\hat{\mathcal{S}}(\hat{\boldsymbol{x}},D)\right|, && (13)\end{aligned}$$

which follows from the same consideration by symmetry. It follows then that

$$|\mathcal{T}_n(P^\ell)| \cdot \left|T_n(Q^\ell)\bigcap \mathcal{S}(\boldsymbol{x},D)\right| = |\mathcal{T}_n(Q^\ell)| \cdot \left|T_n(P^\ell)\bigcap \hat{\mathcal{S}}(\hat{\boldsymbol{x}},D)\right|, \tag{14}$$

or, equivalently,

$$\frac{|T_n(P^\ell)|}{\left|T_n(P^\ell)\bigcap \hat{\mathcal{S}}(\hat{\boldsymbol{x}},D)\right|} = \frac{|T_n(Q^\ell)|}{\left|T_n(Q^\ell)\bigcap \mathcal{S}(\boldsymbol{x},D)\right|}. \tag{15}$$

Now, let Q_*^ℓ be the type of $\hat{\boldsymbol{x}}$ that maximizes $\left|T_n(P^\ell)\bigcap \hat{\mathcal{S}}(\hat{\boldsymbol{x}},D)\right|$. Then, the last equation implies that

$$\frac{|T_n(P^\ell)|}{\max_{\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}^n}\left|T_n(P^\ell)\bigcap \hat{\mathcal{S}}(\hat{\boldsymbol{x}},D)\right|} = \frac{|T_n(Q_*^\ell)|}{\left|T_n(Q_*^\ell)\bigcap \mathcal{S}(\boldsymbol{x},D)\right|}, \quad \forall\, \boldsymbol{x}\in\mathcal{T}_n(P^\ell). \tag{16}$$

This relationship will be used shortly.

Let $P^\ell \in \mathcal{P}_n^\ell$ be given. Any d-semifaithful code must fully cover the type class $\mathcal{T}_n(P^\ell)$ with spheres of radius nD (henceforth, referred to as D-spheres), centered at the various codewords. Let $\hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_M \in \hat{\mathcal{X}}^n$ be M codewords. The number of members of $\mathcal{T}_n(P^\ell)$ that are covered by $\hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_M \in \hat{\mathcal{X}}^n$ is upper bounded as follows.

$$\begin{aligned}
G &= \left|\bigcup_{i=1}^M \left[\mathcal{T}_n(P^\ell)\bigcap \hat{\mathcal{S}}(\hat{\boldsymbol{x}}_i,D)\}\right]\right| \\
&\le \sum_{i=1}^M \left|\mathcal{T}_n(P^\ell)\bigcap \hat{\mathcal{S}}(\hat{\boldsymbol{x}}_i,D)\}\right| \\
&\le M\cdot \max_{\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}^n}\left|\mathcal{T}_n(P^\ell)\bigcap \hat{\mathcal{S}}(\hat{\boldsymbol{x}},D)\}\right|,
\end{aligned} \tag{17}$$

and so, the necessary condition for complete covering, which is $G \ge |\mathcal{T}_n(P^\ell)|$, amounts to

$$\begin{aligned}
M &\ge \frac{|\mathcal{T}_n(P^\ell)|}{\max_{\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}^n}\left|\mathcal{T}_n(P^\ell)\bigcap \hat{\mathcal{B}}(\hat{\boldsymbol{x}},D)\}\right|} \\
&= \frac{|T_n(Q_*^\ell)|}{\left|T_n(Q_*^\ell)\bigcap \mathcal{S}(\boldsymbol{x},D)\right|} \\
&\triangleq M_0,
\end{aligned} \tag{18}$$

where the second line is by (16). Consider now a variable-length code with a codebook of size M. Let $L(\hat{\boldsymbol{x}})$ denote the length (in bits) of the compressed binary string that represents $\hat{\boldsymbol{x}}$. The number of codewords with $L(\hat{\boldsymbol{x}}) \le \log M - \epsilon \log n$ is upper bounded as follows:

$$\begin{aligned}
|\{\hat{\boldsymbol{x}} \in \mathcal{C}_n :\ L(\hat{\boldsymbol{x}}) \le \log M - \epsilon \log n\}| &= \sum_{k=1}^{\log M - \epsilon \log n} |\{\hat{\boldsymbol{x}} :\ L(\hat{\boldsymbol{x}}) = k\}| \\
&\le \sum_{k=1}^{\log M - \epsilon \log n} 2^k \\
&= 2^{\log M - \epsilon \log n + 1} - 1 \\
&< 2n^{-\epsilon} M, \qquad (19)
\end{aligned}$$

where in the first inequality we have used the assumed one-to-one property of the mapping between the reproduction codewords and their variable-length compressed binary representations. It follows then that for at least $M(1 - 2n^{-\epsilon})$ out of the M codewords in $\mathcal{C}^n$ (that is, the vast majority codewords), we have

$$\begin{aligned}
L(\phi_n(\boldsymbol{x})) &\ge \log M - \epsilon \log n \\
&\ge \log M_0 - \epsilon \log n \\
&= -\log\left(\frac{\left|T_n(Q_*^\ell) \bigcap \mathcal{S}(\boldsymbol{x}, D)\right|}{|T_n(Q_*^\ell)|}\right) - \epsilon \log n \\
&= -\log\left[\sum_{\hat{\boldsymbol{x}} \in \mathcal{S}(\boldsymbol{x}, D)} U_{Q_*}(\hat{\boldsymbol{x}})\right] - \epsilon \log n, \qquad (20)
\end{aligned}$$

where U_{Q_*} is the uniform probability distribution across the type class Q_*^ℓ, i.e.,

$$U_{Q_*}(\hat{\boldsymbol{x}}) = \begin{cases} \frac{1}{|\mathcal{T}_n(Q_*^\ell)|} & \hat{\boldsymbol{x}} \in \mathcal{T}_n(Q_*^\ell) \\ 0 & \text{elsewhere} \end{cases} \qquad (21)$$

We now argue that for every $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}^n$

$$U_{Q_*}(\hat{\boldsymbol{x}}) \le \exp_2\{-LZ(\hat{\boldsymbol{x}}) + n\Delta_n(\ell)\}. \qquad (22)$$

For $\hat{\boldsymbol{x}} \notin \mathcal{T}_n(Q_*^\ell)$, this is trivial as the l.h.s. is equal to zero. For $\hat{\boldsymbol{x}} \in \mathcal{T}_n(Q_*^\ell)$, we have the following consideration: Combining eqs. (30) and (32) of [17] together with the inequality [5, p. 17, Lemma 2.3],

$$|\mathcal{T}_n(Q_*^\ell)| \ge \left(\frac{n}{\ell} + 1\right)^{-K^\ell} \cdot 2^{nH_{Q_*}(\hat{X}^\ell)/\ell}, \qquad (23)$$

where

$$H_{Q_*}(\hat{X}^\ell) = -\sum_{b^\ell \in \hat{\mathcal{X}}^\ell} Q_*^\ell(b^\ell) \log Q_*^\ell(b^\ell), \tag{24}$$

we have

$$\begin{aligned}
\log|\mathcal{T}_n(Q_*^\ell)| &\geq c(\hat{\boldsymbol{x}}) \log c(\hat{\boldsymbol{x}}) - n\delta_n(\ell) - \frac{K^\ell}{n} \log\left(\frac{n}{\ell} + 1\right) \\
&\geq LZ(\hat{\boldsymbol{x}}) - n\epsilon(n) - n\delta_n(\ell) - \frac{K^\ell}{n} \log\left(\frac{n}{\ell} + 1\right) \\
&\stackrel{\Delta}{=} LZ(\hat{\boldsymbol{x}}) - n\Delta_n(\ell),
\end{aligned} \tag{25}$$

where

$$\delta_n(\ell) = \frac{\log[4S^2(\ell)] \log K}{(1-\epsilon_n)\log n} + \frac{S^2(\ell)\log[4S^2(\ell)]}{n} + \frac{K^\ell}{n} \log\left(\frac{n}{\ell} + 1\right) + \frac{1}{\ell}, \tag{26}$$

and

$$S(\ell) = \frac{J^{\ell+1} - 1}{J - 1}, \tag{27}$$

and where the second inequality in (25) follows from (3). The last line of (25) is equivalent to (22). It follows then that for at least $M(1 - 2 \cdot n^{-\epsilon})$ out of the M codewords in $\mathcal{C}^n$,

$$\begin{aligned}
L(\phi_n(\boldsymbol{x})) &\geq -\log\left[\sum_{\hat{\boldsymbol{x}} \in \mathcal{S}(\boldsymbol{x}, D)} 2^{-LZ(\hat{\boldsymbol{x}})}\right] - n\Delta_n(\ell) - \epsilon \log n \\
&= -\log\left[\sum_{\hat{\boldsymbol{x}} \in \mathcal{S}(\boldsymbol{x}, D)} \frac{2^{-LZ(\hat{\boldsymbol{x}})}}{\sum_{\hat{\boldsymbol{x}}' \in \hat{\mathcal{X}}^n} 2^{-LZ(\hat{\boldsymbol{x}}')}}\right] - \\
&\quad \log\left(\sum_{\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}^n} 2^{-LZ(\hat{\boldsymbol{x}})}\right) - n\Delta_n(\ell) - \epsilon \log n \\
&\geq -\log(U[\mathcal{S}(\boldsymbol{x}, D)]) - n\Delta_n(\ell) - \epsilon \log n,
\end{aligned} \tag{28}$$

where in the last step we have applied Kraft's inequality to the LZ code-length function. This completes the proof of Theorem 1.

4 The Achievability Theorem

The lower bound of Theorem 1 naturally suggests achievability using the universal distribution, U, for random selection of the various codewords. The basic idea is quite standard and simple: The quantity $U[\mathcal{S}(\boldsymbol{x},D)]$ is the probability that a single randomly chosen reproduction vector, drawn under U, would fall within distance nD from the source vector, $\boldsymbol{x}$. If all reproduction vectors are drawn independently under U, then the typical number of such random selections that it takes before one sees the first one in $\mathcal{S}(\boldsymbol{x},D)$, is of the exponential order of $1/U[\mathcal{S}(\boldsymbol{x},D)]$. Given that the codebook is revealed to both the encoder and decoder, once it has been selected, the encoder merely needs to transmit the index of the first such reproduction vector within the codebook, and the description length of that index can be made essentially as small as $\log\{1/U[\mathcal{S}(\boldsymbol{x},D)]\} = -\log(U[\mathcal{S}(\boldsymbol{x},D))$. We use this simple idea to prove achievability for an arbitrary distortion measure. The proof is very similar to the parallel proof in [14], and it is presented here mainly for completeness.

The achievability theorem is the following.

Theorem 2 *Let $d : \mathcal{X}^n \times \hat{\mathcal{X}}^n \to \mathrm{I\!R}^+$ be an arbitrary distortion function. Then, for every $\epsilon > 0$, there exists a sequence of d-semifaithful, variable-length block codes of block length n, such that for every $\boldsymbol{x} \in \mathcal{X}^n$, the code length for $\boldsymbol{x}$ is upper bounded by*

$$L(\boldsymbol{x}) \le -\log(U[\mathcal{S}(\boldsymbol{x},D)]) + (2+\epsilon)\log n + c + \delta_n, \tag{29}$$

where $c > 0$ is a constant and $\delta_n = O(nJ^n e^{-n^{1+\epsilon}})$.

Proof. The proof is based on the following simple well known fact: Given a source vector $\boldsymbol{x} \in \mathcal{X}^n$ and a codebook, $\mathcal{C}_n$, let $I(\boldsymbol{x})$ denote the index, i, of the first vector, $\hat{\boldsymbol{x}}_i$, such that $d(\boldsymbol{x},\hat{\boldsymbol{x}}_i) \le nD$, namely, $\hat{\boldsymbol{x}}_i \in \mathcal{S}(\boldsymbol{x},D)$. If all reproduction vectors are drawn independently under U, then, for every positive integer, N:

$$\Pr\{I(\boldsymbol{x}) > N\} = (1 - U[\mathcal{S}(\boldsymbol{x},D)])^N = \exp\{N\ln(1-U[\mathcal{S}(\boldsymbol{x},D)]\} \le \exp\{-N \cdot U[\mathcal{S}(\boldsymbol{x},D)]\}, \tag{30}$$

and so, if $N = N_n = e^{\lambda_n}/U[\mathcal{S}(\boldsymbol{x},D)]$, for some arbitrary positive sequence, $\{\lambda_n\}$, that tends to infinity, then

$$\Pr\{I(\boldsymbol{x}) > N_n\} \le \exp\{-e^{\lambda_n}\}. \tag{31}$$

This fact will be used few times in this section.

For later use, we also need the following uniform lower bound to $U[\mathcal{S}(\boldsymbol{x},D)]$: For a given $\boldsymbol{x}$, let $\hat{\boldsymbol{x}}_0 \in \hat{\mathcal{X}}^n$ denote an arbitrary reproduction vector within $\mathcal{S}(\boldsymbol{x},D)$. Then,

$$U[\mathcal{S}(\boldsymbol{x},D)] \quad \ge \quad U(\hat{\boldsymbol{x}}_0) \tag{32}$$

$$
\begin{aligned}
&= \frac{2^{-LZ(\hat{\boldsymbol{x}}_0)}}{\sum_{\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}^n} 2^{-LZ(\hat{\boldsymbol{x}})}} && (33)\\
&\geq 2^{-LZ(\hat{\boldsymbol{x}}_0)}. && (34)
\end{aligned}
$$

Next, observe that $LZ(\hat{\boldsymbol{x}}_0)$ is maximized by the K-ary extension of the counting sequence [33, p. 532], which is defined as follows: For $i = 1, 2, \ldots, m$ (m – positive integer), let $u(i)$ denote the K-ary string of length iK^i that lists, say, in lexicographic order, all the K^i words from $\hat{\mathcal{X}}^i$, and let $\hat{\boldsymbol{x}}_0 = (u(1)u(2)\ldots u(m))$, whose length is

$$
\begin{aligned}
n &= \sum_{i=1}^{m} iK^i\\
&= K \cdot \sum_{i=1}^{m} iK^{i-1}\\
&= K \cdot \frac{\partial}{\partial K}\left(\sum_{i=1}^{m} K^i\right)\\
&= K \cdot \frac{\partial}{\partial K}\left(\frac{K^{m+1}-K}{K-1}\right)\\
&= \frac{K}{(K-1)^2}[mK^{m+1} - (m+1)K^m + 1]. \qquad (35)
\end{aligned}
$$

The LZ incremental parsing of $\hat{\boldsymbol{x}}_0$, which is exactly $(u(1), u(2), \ldots, u(m))$, yields:

$$c(\hat{\boldsymbol{x}}_0) = \sum_{i=1}^{m} K^i = \frac{K^{m+1}-K}{K-1}, \qquad (36)$$

and so, considering eq. (3), it follows that $LZ(\hat{\boldsymbol{x}}_0) \leq (1+\epsilon_n)n\log K$ for some $\epsilon_n \to 0$ as $n \to \infty$.[1] It follows then that

$$U[\mathcal{S}(\boldsymbol{x}, D)] \geq 2^{-n(1+\epsilon_n)\log K}. \qquad (37)$$

Consider now an independent random selection of all reproduction vectors to form a codebook, $\mathcal{C}_n$, of size $M = A^n$ ($A > K$) codewords, $\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \ldots, \hat{\boldsymbol{x}}_M$, according to U. Once the codebook $\mathcal{C}_n = \{\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \ldots, \hat{\boldsymbol{x}}_M\}$ has been drawn, it is revealed to both the encoder and the decoder. Consider next the following encoder. As defined before, let $I(\boldsymbol{x})$ be defined as the index of the first codeword that falls within $\mathcal{S}(\boldsymbol{x}, D)$, but now, with the small modification that if none of the A^n codewords

[1] As an alternative to this upper bound on the LZ code length, one can slightly modify the LZ algorithm as follows: If $LZ(\hat{\boldsymbol{x}}) \leq n\log K$ use the LZ algorithm as usual, otherwise, send $\hat{\boldsymbol{x}}$ uncompressed using $n\log K$ bits. To distinguish between the two modes of operation, append a flag bit to indicate whether or not the data is LZ-compressed. The modified code-length would then be $LZ'(\hat{\boldsymbol{x}}) = \min\{LZ(\hat{\boldsymbol{x}}), n\log K\} + 1$. Now, replace $LZ(\hat{\boldsymbol{x}})$ by $LZ'(\hat{\boldsymbol{x}})$ in all places throughout this paper, including the definition of U.

fall in $\mathcal{S}(\boldsymbol{x}, D)$, then we define $I(\boldsymbol{x}) = A^n$ nevertheless (and then the encoding fails). Next, we define the following probability distribution over the positive integers, $1, 2, \ldots, A^n$:

$$u[i] = \frac{1/i}{\sum_{k=1}^{A^n} 1/k}, \qquad i = 1, 2, \ldots, A^n. \tag{38}$$

Given $\boldsymbol{x}$, the encoder finds $I(\boldsymbol{x})$ and encodes it using a variable-rate lossless code with the length function (in bits, and ignoring the integer length constraint),

$$\begin{aligned} L(\boldsymbol{x}) &= -\log u[I(\boldsymbol{x})] \\ &\le \log I(\boldsymbol{x}) + \log\left(\sum_{k=1}^{A^n} \frac{1}{k}\right) \\ &\le \log I(\boldsymbol{x}) + \log(\ln A^n + 1) \\ &= \log I(\boldsymbol{x}) + \log(n \ln A + 1) \\ &\le \log I(\boldsymbol{x}) + \log n + c, \end{aligned} \tag{39}$$

where $c = \log(\ln A + 1)$. It follows that the expected codeword length for $\boldsymbol{x} \in \mathcal{X}^n$ (w.r.t. the randomness of the code) is upper bounded by:

$$\begin{aligned} \boldsymbol{E}\{L(\boldsymbol{x})\} &\le \boldsymbol{E}\{\log I(\boldsymbol{x})\} + \log n + c \\ &\le \log \boldsymbol{E}\{I(\boldsymbol{x})\} + \log n + c \\ &= \log\left(\sum_{k=1}^{A^n} k \cdot (1 - U[\mathcal{S}(\boldsymbol{x}, D)])^{k-1} \cdot U[\mathcal{S}(\boldsymbol{x}, D)] + A^n \cdot (1 - U[\mathcal{S}(\boldsymbol{x}, D)])^{A^n}\right) + \log n + c \\ &= \log\left(\sum_{k=1}^{\infty} \min\{k, A^n\} \cdot (1 - U[\mathcal{S}(\boldsymbol{x}, D)])^{k-1} \cdot U[\mathcal{S}(\boldsymbol{x}, D)]\right) + \log n + c \\ &\le \log\left\{\sum_{k=1}^{\infty} k \cdot (1 - U[\mathcal{S}(\boldsymbol{x}, D)])^{k-1} \cdot U[\mathcal{S}(\boldsymbol{x}, D)]\right\} + \log n + c \\ &= \log\left(\frac{1}{U[\mathcal{S}(\boldsymbol{x}, D)]}\right) + \log n + c, \end{aligned} \tag{40}$$

and we denote

$$L^+(\boldsymbol{x}) \stackrel{\Delta}{=} \log\left(\frac{1}{U[\mathcal{S}(\boldsymbol{x}, D)]}\right) + \log n + c. \tag{41}$$

Consider now the quantity

$$\begin{aligned} E_n &\stackrel{\Delta}{=} \boldsymbol{E}\bigg\{\max\bigg(\max_{\boldsymbol{x} \in \mathcal{X}^n} \mathcal{I}\{d(\boldsymbol{x}, \hat{\boldsymbol{X}}) > nD\}, \\ &\qquad \left[\max_{\boldsymbol{x} \in \mathcal{X}^n} \left(L(\boldsymbol{x}) - L^+(\boldsymbol{x}) - (1 + \epsilon)\log n\right)\right]_+\bigg)\bigg\}, \end{aligned} \tag{42}$$

where the expectation is w.r.t. the randomness of the code, $\mathcal{C}_n$. If E_n can be upper bounded by δ_n, which tends to zero as $n \to \infty$, this will imply that there must exist a code for which both

$$\max_{\boldsymbol{x}\in\mathcal{X}^n} \mathcal{I}\{d(\boldsymbol{x},\hat{\boldsymbol{x}}) > nD\} \le \delta_n \tag{43}$$

and

$$\max_{\boldsymbol{x}\in\mathcal{X}^n} (L(\boldsymbol{x}) - L^+(\boldsymbol{x}) - (1+\epsilon)\log n) \le \delta_n \tag{44}$$

at the same time. Observe that since the left-hand side of (43) is either zero or one, then if we know that it must be less than $\delta_n \to 0$, for some codebook, $\mathcal{C}_n$, it means that it must vanish as soon as n is large enough such that $\delta_n < 1$, namely, $d(\boldsymbol{x},\hat{\boldsymbol{x}}) \le nD$ for all $\boldsymbol{x}$, in other words, the code is d-semifaithful. Also, by (44), for the same codebook, we must have

$$L(\boldsymbol{x}) \le L^+(\boldsymbol{x}) + (1+\epsilon)\log n + \delta_n \qquad \boldsymbol{x} \in \mathcal{X}^n, \tag{45}$$

and δ_n adds a negligible redundancy term.

To prove that $E_n \to 0$, we first use the simple fact that the maximum of two non-negative numbers is upper bounded by their sum, i.e.,

$$\begin{aligned} E_n &\le \boldsymbol{E}\left\{\max_{\boldsymbol{x}\in\mathcal{X}^n} \mathcal{I}\{d(\boldsymbol{x},\hat{\boldsymbol{X}}) > nD\}\right\} + \\ &\quad \boldsymbol{E}\left\{\left[\max_{\boldsymbol{x}\in\mathcal{X}^n} (L(\boldsymbol{x}) - L^+(\boldsymbol{x}) - (1+\epsilon)\log n)\right]_+\right\}, \end{aligned} \tag{46}$$

and therefore, it is sufficient to prove that each one of these terms tends to zero. As for the first term, we have:

$$\begin{aligned} \boldsymbol{E}\left\{\max_{\boldsymbol{x}\in\mathcal{X}^n} \mathcal{I}\{d(\boldsymbol{x},\hat{\boldsymbol{X}}) > nD\}\right\} &\le \boldsymbol{E}\left\{\sum_{\boldsymbol{x}\in\mathcal{X}^n} \mathcal{I}\{d(\boldsymbol{x},\hat{\boldsymbol{X}}) > nD\}\right\} \\ &= \sum_{\boldsymbol{x}\in\mathcal{X}^n} \boldsymbol{E}\left\{\mathcal{I}\{d(\boldsymbol{x},\hat{\boldsymbol{X}}) > nD\}\right\} \\ &= \sum_{\boldsymbol{x}\in\mathcal{X}^n} \Pr\{d(\boldsymbol{x},\hat{\boldsymbol{X}}) > nD\} \\ &= \sum_{\boldsymbol{x}\in\mathcal{X}^n} (1 - U[\mathcal{S}(\boldsymbol{x},D)])^{A^n} \\ &\le \sum_{\boldsymbol{x}\in\mathcal{X}^n} \exp\{-A^n U[\mathcal{S}(\boldsymbol{x},D)]\} \\ &\overset{\text{(a)}}{\le} \sum_{\boldsymbol{x}\in\mathcal{X}_n} \exp\{-\exp\{n[\ln A - (1+\epsilon_n)\ln K]\}\} \\ &\le J^n \exp(-\exp\{n[\ln A - (1+\epsilon_n)\ln K]\}\}, \end{aligned} \tag{47}$$

where in (a) we have used (37). This quantity decays double-exponentially rapidly as $n \to \infty$ since we have assumed $A > K$.

As for the second term of (46), we have:

$$\begin{aligned}
& \boldsymbol{E}\left\{\left[\max_{\boldsymbol{x}\in\mathcal{X}^n}\left(L(\boldsymbol{x})-L^+(\boldsymbol{x})-(1+\epsilon)\log n\right)\right]_+\right\} \\
\overset{\text{(a)}}{\le} \; & \boldsymbol{E}\left\{\left[\max_{\boldsymbol{x}\in\mathcal{X}^n}\left(\log I(\boldsymbol{x})-\log\frac{1}{U[\mathcal{S}(\boldsymbol{x},D)]}-(1+\epsilon)\log n\right)\right]_+\right\} \\
= \; & \int_0^\infty \Pr\left\{\max_{\boldsymbol{x}\in\mathcal{X}^n}\left[\log I(\boldsymbol{x})-\log\frac{1}{U[\mathcal{S}(\boldsymbol{x},D)]}-(1+\epsilon)\log n)\right]\ge s\right\}\mathrm{d}s \\
= \; & \int_0^{n\log A} \Pr\left\{\max_{\boldsymbol{x}\in\mathcal{X}^n}\left[\log I(\boldsymbol{x})-\log\frac{1}{U[\mathcal{S}(\boldsymbol{x},D)]}-(1+\epsilon)\log n\right]\ge s\right\}\mathrm{d}s \\
= \; & \int_0^{n\log A} \Pr\left[\bigcup_{\boldsymbol{x}\in\mathcal{X}^n}\left\{I(\boldsymbol{x})\ge\frac{2^{(1+\epsilon)\log n+s}}{U[\mathcal{S}(\boldsymbol{x},D)]}\right\}\right]\mathrm{d}s \\
\le \; & \sum_{\boldsymbol{x}\in\mathcal{X}^n}\int_0^{n\log A}\Pr\left\{I(\boldsymbol{x})\ge\frac{2^{(1+\epsilon)\log n+s}}{U[\mathcal{S}(\boldsymbol{x},D)]}\right\}\mathrm{d}s \\
\overset{\text{(b)}}{\le} \; & \sum_{\boldsymbol{x}\in\mathcal{X}^n}\int_0^{n\log A}\exp\{-2^s n^{1+\epsilon}\}\mathrm{d}s \\
\le \; & J^n\cdot(n\log A)\cdot\exp\{-n^{1+\epsilon}\}, \qquad (48)
\end{aligned}$$

where in (a) we have used (39) and (41), and in (b) we have used (31). The right-most side of this chain of inequalities clearly decays as well when n grows without bound. This completes the proof.

5 Summary and Discussion

By deriving asymptotically matching upper and lower bounds, we have established the quantity $-\frac{1}{n}\log(U[\mathcal{S}(\boldsymbol{x},D)])$ as having the significance of an empirical rate distortion function for individual sequences. While this quantity is not easy to calculate for large n, the operative meaning of our results is that we propose a universal ensemble for rate-distortion coding. According to this ensemble, the codewords are drawn independently under the probability distribution that is proportional to $2^{-LZ(\hat{\boldsymbol{x}})}$.

There are several observations, insights and perspectives that should be addressed.

Relation to earlier converse bounds. The converse bound is given in terms of the probability of a sphere of radius nD around the source vector $\boldsymbol{x}$, under the universal distribution, U, defined in (4). This is intimately related to a converse result due to Kontoyiannis and Zhang [11, Theorem 1, part i)], which states that for any d-semifaithful code, there exists a probability distribution Q on $\hat{\mathcal{X}}^n$ such that $L(\boldsymbol{x}) \geq -\log(Q[\mathcal{S}(\boldsymbol{x}, D)])$ for all $\boldsymbol{x}$ (see also [10]). Here, upon giving up any claims on a minority of the codewords pertaining to a given type class, we derived a lower bound of essentially the same form with the benefit of specifying a concrete choice of the distribution Q, i.e., we propose $Q = U$, the universal distribution (unlike the distribution in [11, Section III.A], which is proportional to $2^{-L(\hat{\boldsymbol{x}})}$ across the codebook).

Interpretation of the main term of the bound. Since $LZ(\hat{\boldsymbol{x}})$ is essentially bounded by a linear function of n (see (37)), we can approximate the main term as follows:

$$\begin{aligned}
-\log(U[\mathcal{S}(\boldsymbol{x}, D)]) &\leq -\log\left(\sum_{\hat{\boldsymbol{x}}\in\mathcal{S}(\boldsymbol{x},D)} 2^{-LZ(\hat{\boldsymbol{x}})}\right) \\
&= -\log\left(\sum_{L\geq 1} 2^{-L}\cdot\left|\{\hat{\boldsymbol{x}}:\ LZ(\hat{\boldsymbol{x}}) = L\}\bigcap\mathcal{S}(\boldsymbol{x}, D)\right|\right) \\
&\approx \min_{L\geq 1}\left\{L - \log\left|\{\hat{\boldsymbol{x}}:\ LZ(\hat{\boldsymbol{x}}) = L\}\bigcap\mathcal{S}(\boldsymbol{x}, D)\right|\right\}. \qquad (49)
\end{aligned}$$

This expression, when normalized by n, can be viewed as a certain extension of the rate distortion function, from the memoryless case to the general case, in the following sense: For a memoryless source P, the rate-distortion function has the following representation, which is parallel to (49):

$$R(D) = \min_{P_{\hat{X}}}\left[H(\hat{X}) - \max_{\{P_{X|\hat{X}}:\ \boldsymbol{E}d(X,\hat{X})\leq D,\ P_X = P\}} H(\hat{X}|X)\right], \qquad (50)$$

where the maximum over the empty set is understood to be $-\infty$. Indeed, if we replace U by the the uniform distribution across the first-order type pertaining to the optimal $P_{\hat{X}}$, this is the corresponding single-letter expression of $-\log(P_{\hat{X}}[\mathcal{S}(\boldsymbol{x}, D)])$ that is obtained using the method of types [5].

Comparing to the LZ description length of the most compressible $\hat{\boldsymbol{x}} \in \mathcal{S}(\boldsymbol{x}, D)$. Since our achievable bound involves LZ compression, it is interesting to compare it to the conceptually simple coding scheme that encodes $\boldsymbol{x}$ by the vector $\hat{\boldsymbol{x}}$ that minimizes $LZ(\hat{\boldsymbol{x}})$ within $\mathcal{S}(\boldsymbol{x},D)$. Consider the following chain of equalities and inequalities:

$$\begin{aligned}\min_{\hat{\boldsymbol{x}}\in\mathcal{S}(\boldsymbol{x},D)} LZ(\hat{\boldsymbol{x}}) &= -\log\left(\max_{\hat{\boldsymbol{x}}\in\mathcal{S}(\boldsymbol{x},D)} 2^{-LZ(\hat{\boldsymbol{x}})}\right)\\ &\geq -\log\left(\sum_{\hat{\boldsymbol{x}}\in\mathcal{S}(\boldsymbol{x},D)} 2^{-LZ(\hat{\boldsymbol{x}})}\right)\\ &\geq -\log\left(\sum_{\hat{\boldsymbol{x}}\in\mathcal{S}(\boldsymbol{x},D)} \frac{2^{-LZ(\hat{\boldsymbol{x}})}}{\sum_{\hat{\boldsymbol{x}}'\in\hat{\mathcal{X}}^n} 2^{-LZ(\hat{\boldsymbol{x}}')}}\right)\\ &= -\log(U[\mathcal{S}(\boldsymbol{x},D)]), \end{aligned} \tag{51}$$

which means that the performance of our proposed scheme is never worse (and conceivably, often much better) than that of selecting the vector $\hat{\boldsymbol{x}}$ with the smallest $LZ(\hat{\boldsymbol{x}})$ among all reproduction vectors in $\mathcal{S}(\boldsymbol{x},D)$. The reason for the superiority of the proposed scheme is that it takes advantage of the fact that $\hat{\boldsymbol{x}}$ cannot be any vector in $\hat{\mathcal{X}}^n$, but it must be a member of the codebook, $\mathcal{C}_n$, i.e., one of the possible outputs of a vector quantizer. On the other hand, in view of [33], $\min_{\hat{\boldsymbol{x}}\in\mathcal{S}(\boldsymbol{x},D)} LZ(\hat{\boldsymbol{x}})$ is essentially achievable upon compressing the output of a certain reproduction encoder (or vector quantizer) using a finite–state encoder, but a finite-state machine does not have enough memory resources to take advantage of the fact that vectors outside $\mathcal{C}_n$ cannot be encountered by the encoder. Another interesting comparison between the two schemes is in terms of computational complexity. While in our scheme, the encoder has to carry out typically about $1/U[\mathcal{S}(\boldsymbol{x},D)]$ distortion calculations before finding the first $\hat{\boldsymbol{x}}\in\mathcal{S}(\boldsymbol{x},D)$, in the alternative scheme the number of calculations is $|\mathcal{S}(\boldsymbol{x},D)|$. The former is decreasing function of D, whereas the latter is an increasing function of D. Therefore, in terms of computational complexity, the preference between the two schemes might depend on D. Specifically, for an additive distortion measure, it is easy to see that

$$\frac{1}{U[\mathcal{S}(\boldsymbol{x},D)]} \overset{\cdot}{\leq} \exp_2\{nR(D,P^1_{\boldsymbol{x}})\} \tag{52}$$

and, by the method of types [5]:

$$|\mathcal{S}(\boldsymbol{x},D)| \doteq \exp_2\{nE(D,P^1_{\boldsymbol{x}}))\} \stackrel{\Delta}{=} \exp_2[\max\{H(\hat{X}|X),\ \boldsymbol{E}d(X,\hat{X})\leq D,\ P_X=P^1_{\boldsymbol{x}}\}]. \tag{53}$$

Therefore, whenever D is large enough such that $R(D,P^1_{\boldsymbol{x}}) < E(D,P^1_{\boldsymbol{x}}))$, it is guaranteed that the coding scheme proposed here is computationally less demanding than the alternative scheme of minimizing $LZ(\hat{\boldsymbol{x}})$ across $\mathcal{S}(\boldsymbol{x}, D)$.

Implementation of the random coding distribution. The universal random coding distribution is not difficult to implement. One way to do this is by feeding the LZ decoder with a sequence of purely random bits (fair coin tosses) until we have obtained n symbols at the decoder output. The details can be found in [20].

Universality w.r.t. the distortion measure. As mentioned in the Introduction, in [12], [13] and [14], there are results on the existence of rate-distortion codes that are universal, not only in terms of the source, but also in the sense of the distortion measure. Since the proof of our achievability scheme is very similar to that of [14], it is possible to extend the achievability proof here too, so as to make our code distortion-universal for a wide class of distortion measures. This can be carried out by redefining E_n to include maximization of both terms over a dense grid of distortion functions, as was done in [14]. We opted not to include this in the present paper since it is straightforward, given the results we already have here and in [14].

References

[1] E. Arikan and N. Merhav, “Guessing subject to distortion,” *IEEE Trans. Inform. Theory*, vol. 44, no. 3, pp. 1041–1056, May 1998.

[2] T. Berger, *Rate Distortion Theory - A Mathematical Basis for Data Compression*, Prentice-Hall Inc., Englewood Cliffs, N.J., 1971.

[3] A. Cohen and N. Merhav, “Universal randomized guessing subjected to distortion,” *IEEE Trans. Inform. Theory*, vol. 68, no. 12, pp. 7714–7734, December 2022.

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

[5] I. Csiszár and J. Körner, *Information Theory - Coding Theorems for Discrete Memoryless Systems*, Second Edition, Cambridge University Press, Cambridge, UK, 2011.

[6] L. D. Davisson, ‘Universal noiseless coding,” *IEEE Trans. Inform. Theory*, vol. IT–29, no. 6, pp. 783–795, November 1973.

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

[8] R. G. Gallager, “Source coding with side information and universal coding,” LIDS-P-937, M.I.T., 1976.

[9] R. M. Gray, *Source Coding Theory*, Kluwer Academic Publishers, Boston, 1990.

[10] I. Kontoyiannis, “Pointwise redundancy in lossy data compression and universal lossy data compression,” *IEEE Trans. Inform. Theory*, vol. 46, no. 1, pp. 136-152, January 2000.

[11] I. Kontoyiannis and J. Zhang, “Arbitrary source models and Bayesian codebooks in rate-distortion theory,” *IEEE Trans. Inform. Theory*, vol. 48, no. 8, pp. 2276–2290, August 2002.

[12] A. Mahmood and A. B. Wagner, “Lossy compression with universal distortion,” `https://arxiv.org/pdf/2110.07022.pdf` February 9, 2022.

[13] A. Mahmood and A. B. Wagner, "Minimax rate-distortion," `https://arxiv.org/pdf/2202.04481.pdf` February 9, 2022.

[14] N. Merhav, "D-semifaithful codes that are universal over both memoryless sources and distortion measures," submitted for publication. Also, available on-line at: `http://arxiv.org/pdf/2203.03305.pdf`

[15] N. Merhav, "A comment on 'A rate of convergence result for a universal $d-$semifaithful code'," *IEEE Trans. Inform. Theory*, vol. 41, no. 4, pp. 1200-1202, July 1995.

[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, "Guessing individual sequences: generating randomized guesses using finite-state machines," *IEEE Trans. Inform. Theory*, vol. 66, no. 5, pp. 2912–2920, May 2020.

[18] N. Merhav, "Encoding individual source sequences for the wiretap channel," *Entropy*, 23(12) 1694, December 17, 2021.

[19] N. Merhav, "Finite-state source-channel coding for individual source sequences with source side information at the decoder," *IEEE Trans. Inform. Theory*, vol. 68, no. 3, pp. 1532–1544, March 2022.

[20] N. Merhav and A. Cohen, "Universal randomized guessing with application to asynchronous decentralized brute–force attacks," *IEEE Trans. Inform. Theory*, vol. 66, no. 1, pp. 114–129, January 2020.

[21] N. Merhav and M. Feder, "A strong version of the redundancy–capacity theorem of universal coding," *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 714-722, May 1995.

[22] 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.

[23] M. J. Weinberger, N. Merhav, and M. Feder, "Optimal sequential probability assignment for individual sequences," *IEEE Trans. Inform. Theory*, vol. 40, no. 2, pp. 384–396, March 1994.

[24] D. S. Ornstein and P. C. Shields, "Universal almost sure data compression," *Ann. Probab.*, vol. 18, no. 2, pp. 441–452, 1990.

[25] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Transactions on Information Theory*, vol. IT–30, no. 4, pp. 629–636, July 1984.

[26] J. F. Silva and P. Piantanida, "On universal d-semifaithful coding for memoryless sources with infinite alphabets," *IEEE Transactions on Information Theory*, vol. 68, no. 4, pp. 2782–2800, April 2022.

[27] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*, McGraw-Hill Inc., New York, 1979.

[28] B. Yu and T. Speed, "A rate of convergence result for a universal d-semifaithful code," *IEEE Trans. Inform. Theory*, vol. 39, no. 3, pp. 813–820, May 1993.

[29] Z. Zhang, E.-h. Yang, and V. Wei, "The redundancy of source coding with a fidelity criterion. I. known statistics," *IEEE Trans. Inform. Theory*, vol. 43, no. 1, pp. 71–91, January 1997.

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

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

[32] 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.

[33] 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.