D-Semifaithful Codes That are Universal Over Both Memoryless Sources and Distortion Measures

Neri Merhav, Life Fellow, IEEE

Abstract—We prove the existence of codebooks for d-semifaithful lossy compression that are simultaneously universal with respect to both the class of finite-alphabet memoryless sources and the class of all bounded additive rational distortion measures. By applying independent random selection of the codewords according to a mixture of all memoryless sources, we achieve redundancy rates that are within \( O(\log n/n) \) close to the empirical rate-distortion function of every given source vector with respect to every bounded, rational distortion measure.

Index Terms—Lossy compression, rate-distortion theory, universal coding, random coding.

I. INTRODUCTION

We consider the classical problem of lossy compression for finite-alphabet memoryless sources with respect to (w.r.t.) a fidelity criterion defined by an additive distortion measure \([4], [5, \text{Chap. 10}], [8, \text{Chap. 9}], [12], [8, \text{Chaps. 7}]\). More specifically, our focus is on d-semifaithful codes, i.e., variable-length codes that meet a given distortion constraint for each and every source sequence (and not only on the average). As is very well known \([4]\), the rate-distortion function characterizes the least achievable expected coding rate for a given memoryless source and distortion measure.

Motivated by the consideration that the source statistics are seldom known in practice, many research efforts, throughout the years, have been devoted to the quest for universal codes, namely, codes that are independent of the unknown memoryless source, but nevertheless, achieve the fundamental limits (entropy, in the lossless case, and the rate-distortion function, in the lossy case) asymptotically, for long blocks. In the next two paragraphs we provide a very brief account of a few earlier works, which are directly relevant to the current work. This is by no means a comprehensive literature review.

In the lossless case, the theory and practice of universal source coding is fairly mature. Early work by Davisson \([9]\) on the expected redundancy of universal codes contributed the notions of weak (maximin) and strong (minimax) universality, and established their relation to the capacity of an auxiliary channel whose input is the unknown index (or parameter vector) of the source in the class and whose output is the data to be compressed (see also \([11]\)). In the case of a parametric class of sources, the optimal universal-coding redundancy is well-known to be \( \frac{k \log n}{2n} \) plus higher order terms, where \( k \) is the dimension of the parameter vector (or the number of degrees of freedom) and \( n \) is the block-length. One of the pivotal ideas that arises from Davisson’s theory is to construct a universal code whose length function is given by the negative logarithm of the probability of the data vector w.r.t. a Bayesian mixture (with a carefully chosen weight function) of all sources in the class.

Later studies have improved the aforementioned results by \( \frac{1}{2} \log n \) factor \( \frac{k \log n}{2n} \) being replaced by the aforementioned channel capacity. In \([25]\), a converse to the universal-coding theorem, asserting that for large \( n \), no universal code can achieve redundancy smaller than \( (1 - \epsilon) \frac{k \log n}{2n} \), for most sources in the class, i.e., except perhaps for sources that belong to a vanishingly small volume in the parameter space, no matter how small \( \epsilon > 0 \) may be. Merhav and Feder \([22]\) have extended this finding to a general class of sources (not necessarily parametric) with the factor \( \frac{1}{2} \log n \) being replaced by the aforementioned channel capacity. In \([29]\), Weinberger, Merhav and Feder provided an individual-sequence (pointwise) analogue of Rissanen’s result, where the same redundancy term, \( \frac{k \log n}{2n} \), continues to play a role, as the best achievable redundancy, beyond the empirical entropy of the source vector, which applies to most source sequences in each type class, for almost all type classes. Later studies have improved the aforementioned results by using more refined analysis techniques and also extended them in several directions, including their relation to universal prediction theory, see, e.g., \([3]\) and \([21]\).

In the broader scope of universal lossy coding, the parallel theory is less developed and the fundamental limits are not as well understood as in the lossless case. Our focus in this work (and hence also in this brief review) will be on d-semifaithful codes \([24]\), which in the probabilistic case, means codes that comply with the distortion constraint almost surely. As was shown by Zhang et al. \([31]\), in contrast to the lossless case, in the lossy case, even if the source statistics are fully known, it is not possible to achieve expected redundancy smaller than \( \frac{\log n}{2n} \) (see also \([18]\)), but it is possible to achieve \( \log n \).

When the source is unknown, the cost of universality is in having a larger coefficient in front of the \( \frac{\log n}{2n} \) redundancy term. Accordingly, Yu and Speed \([31]\) have proved weak universality with a coefficient that depends on the alphabet sizes of both the source and the reconstruction (see also \([26]\)).
Ournstein and Shields [24] have studied universal \( d \)-semifaithful coding for stationary and ergodic sources under the Hamming distortion measure, and have proved almost-sure convergence to the rate-distortion function. Kontoyiannis [13], in a study of almost-sure redundancy rates, has established several interesting results: The first is a one-sided central limit theorem (CLT), with a \( O(1/\sqrt{n}) \) redundancy term, whose coefficient is characterized as an asymptotically Gaussian random variable with a certain fixed variance. The second is referred to as a law of iterated logarithm (LIL) with a redundancy term proportional to \( \sqrt{\frac{\log \log n}{n}} \) infinitely often almost surely. One of the intriguing findings in [13] is that universality has no cost under these performance measures. In [14], Kontoyiannis and Zhang based many of their results on their finding 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. In that work, they also propose the idea of random coding under a probability distribution that is formed by a Bayesian mixture of all probability distributions in a certain class, like the class of memoryless sources over the reproduction alphabet. More recently, Mahmood and Wagner [16], [17] analyzed \( d \)-semifaithful codes that are strongly universal w.r.t. both the source and the distortion measure (more details will be provided in the sequel). The universal expected redundancy rates in [16] are all of the order of \( \frac{\log n}{n} \) with various multiplicative constants.

In a recent work coauthored with Cohen and Merhav [7] (which is a further development over [2] and [20]), we considered the intimately related problem of universal guessing subject to a fidelity criterion, where the universality takes place in a multitude of dimensions. One of those dimensions is the distortion measure. In this paper, the ideas of [7], as well as those of [14] and [16], are harnessed and considerably refined to demonstrate the existence of \( d \)-semifaithful codes, which are not only universal w.r.t. the source statistics, but also universal w.r.t. the class of all bounded, rational, single-letter distortion measures. In other words, the same universal codebook is completely flexible to be used, not only for one given distortion measure, but for all bounded rational distortion measures, on the top of its universality property for all memoryless sources of a given alphabet, as before. This means that it is enough that the distortion measure would be specified once a source vector has to be actually encoded, and not necessarily before the codebook is constructed. In other words, the idea is that the codebook is universal, but the actual encoding function that is based on that codebook depends on the distortion measure (more details will be provided in the sequel).

The motivations for studying the scenario of universality w.r.t. the distortion measure are as described in [16]: One motivation is that we would like modern compression systems to allow some flexibility to comply with the needs of various users, that may have different preferences with regard to distortion criteria, for example, in images and video, different distortion measures can be used to yield different effects, depending on the sensitivity of the vision of the individual user, for example, balancing between errors in high and low frequencies (if coding is carried out in the frequency domain), and so on. The same kind of rationale applies to audio signals as well. Another motivation is associated with non-linear transform coding, and the details can be found in [16].

As mentioned earlier, Mahmood and Wagner have also provided very interesting results along the very same line [16], [17]. In [16], they proposed three strongly universal coding schemes. The first two are based on unions of codebooks associated with distortion measures that belong to a fine grid in the space of all bounded distortion matrices. The third scheme is based on the notion of the in Vapnik-Chervonenkis (VC) dimension [27]. All three coding schemes achieve rate redundancies that are asymptotically proportional to \( \frac{\log n}{n} \) for blocks of length \( n \), but they differ in the constants of proportionality. In [16] and [17], the focus is more towards strong universality and minimax properties of universal codes. Accordingly, several coding theorems are provided in [17], but the uniformity comes at the inevitable price of a slowdown in the decay of the rate redundancies.

Our approach is conceptually simpler than those of [16] and [17], and we show that smaller redundancies are achievable. Our redundancy results are pointwise (for each and every source sequence), unlike the expected redundancy analysis in Mahmood and Wagner’s works. We also provide a pointwise converse theorem. However, for the sake of fairness, it must be pointed out that in contrast to [17], we make no claims concerning uniformity of convergence.

At the heart of our analysis, we use the saddle-point method to derive the probability that a randomly selected codeword would fall within distortion \( nD \) away from a source sequence of a given type class. The resulting bound is asymptotically tight in the sense that, it does not only have the correct exponential behavior, but moreover, the ratio between the bound and the exact probability tends to unity as the block length \( n \) grows without bound.

The outline of the remaining part of this paper is as follows. In Section II, we establish the notation and formalize the problem. In Section III, we state and prove a lemma that provides an asymptotically tight evaluation of the probability that a random codeword happens to lie within distortion \( nD \) away from the source vector. In Section IV, we state and prove the main coding theorem concerning the universality. In Section V, we prove a pointwise lower bound (converse result). Finally, in Section VI, we summarize this work and provide an outlook for possible future work.

II. NOTATION AND PROBLEM SETTING

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. For example, the random vector \( \mathbf{X} = (X_1, \ldots, X_n) \), \( n \)-positive integer may take a specific vector value \( \mathbf{x} = (x_1, \ldots, x_n) \) in \( \mathcal{X}^n \), the \( n \)–th order Cartesian power of \( \mathcal{X} \), which is the alphabet of each component of this vector. Sources and
As is well known, the rate-distortion coding theorem asserts that for a given memoryless source $P$ and distortion measure $d$, there exist $d$-semifaihtiful codes, $(\phi_n, \psi_n)$, w.r.t. distortion level $D$, whose average coding rate $R$ is arbitrarily close to
\[
R_d(D, P) \triangleq \min_{(P_X | X) : E(d(X, \hat{X})) \leq D} I(X; \hat{X}),
\]
for all sufficiently large $n$. On the other hand, the converse theorem asserts that there are no $d$-semifaihtiful codes w.r.t. distortion level $D$ with $R < R_d(D, P)$.

The following Lagrange-dual representation of $R_d(D, P)$ (in nats per source symbol) is well known (see, e.g., [12, p. 90, Corollary 4.2.3]):
\[
R_d(D, P) = \sup_{s \geq 0} \min_{Q \in \mathcal{P}_n} \{ -\sum_{x \in \mathcal{X}} P(x) \times \ln \left[ \sum_{\hat{x} \in \hat{X}} Q(\hat{x}) e^{-sd(x, \hat{x})} \right] - sD \} = \min_{s \geq 0} \sup_{Q} \left\{ -\sum_{x \in \mathcal{X}} P(x) \times \ln \left[ \sum_{\hat{x} \in \hat{X}} Q(\hat{x}) e^{-sd(x, \hat{x})} \right] - sD \right\},
\]
where minimization is over all probability assignments, $Q = \{Q(\hat{x}), \hat{x} \in \hat{X}\}$, across the reproduction alphabet, $\hat{X}$. Here, the second equality holds since the function,
\[
F(s, Q) \triangleq -\sum_{x \in \mathcal{X}} P(x) \ln \left[ \sum_{\hat{x} \in \hat{X}} Q(\hat{x}) e^{-sd(x, \hat{x})} \right] - sD
\]
is convex in $Q$ and concave in $s$.

Our objective is to prove that there exists a sequence of codebooks with a double universality property in the sense that they are simultaneously $d$-semifaihtiful w.r.t. $D$. For every rational distortion measure $d$ that satisfies requirements of Section II, and, at the same time, their normalized code-length functions are arbitrarily close to $R_d(D, \hat{P}_x)$ for all $x \in \mathcal{X}^n$ when $n$ is sufficiently large $n$. We will also focus on the pointwise achievable redundancy as a function of $n$, i.e., on the quantity $L_d(x)/n - R_d(D, \hat{P}_x)$ for each and every $x$, where $L_d(x)$ is the length of the compressed representation of $x$ when the distortion measure is $d$.

It should be emphasized that in our scenario, the double universality is a property of the codebook, $C_n$, and the decoding function, $\psi_n$, but not the encoding function, $\phi_n$. The encoding function, $\phi_n$, is universal over the types, $\{\hat{P}_x\}$, but it depends on the distortion measure $d$, as it selects the first reproduction vector, $\hat{x} \in C_n$, that meets the distortion constraint w.r.t. a given $d$. With regard to the codebook itself, the important double universality property pertains to the order in which the codewords appear in the codebook and not just to the members of the codebook (similarly as in the guessing problem). This is because the index of the first codeword that meets the distortion constraint is losslessly encoded using a code-length function that increases (essentially logarithmically) with this index (see Section IV). More precisely, the characterization of the system is as follows. The encoding mechanism consists of

\[
\text{The need for the subscript } d \text{ will become clear in the sequel.}
\]
a cascade of two units: a reproduction encoder (or, a vector quantizer), which maps $x$ into $\hat{x}$, and an entropy coder, which losslessly compresses the index of $\hat{x}$ (within the codebook) into a variable-length bit string. The decoder, $\psi_n$, implements just the inverse operation of the entropy coder: it maps the compressed binary string back into $\hat{x}$. The reproduction encoder is defined by a codebook where the codewords are ordered in a certain manner, and a reproduction rule (or, a quantization rule). The codebook, along with its order, and the decoder are universal in both $P_\mathcal{X}$ and $d$. The quantization rule, on the other hand, is universal in $P_\mathcal{X}$, but it depends on $d$.

Our main performance bound (Theorem 1) is pointwise, for each and every source vector $x$. Roughly speaking, the achievable code-length performance for a given source vector $x$ is characterized in terms of its empirical rate-distortion function, i.e., the rate-distortion function associated with the empirical distribution of $x$, $\hat{P}_\mathcal{X}$. In principle, the bound applies within a small term that tends to zero as $n \to \infty$, but the subtlety is that the convergence to zero may not be uniform in $\{x\}$. Fortunately, it depends on $x$ only via the type class (or the empirical distribution) of $x$, which means that there is clearly uniformity within each and every type class. However, to state an asymptotic achievability result of a universal code, we have to assert that there exists a sequence of codes that is simultaneously good for every type class. There are two possible approaches to this end. The first is to let the probability distribution, $P$, pertaining to a type, be any rational probability distribution, i.e., $P \in \bigcup_{k \geq 1} P_k$, and then let the index $n$ take values only in the subsequence of which $P \in P_n$. The second approach is to allow every $n$, but to refer to a sequence of empirical distributions that tends to some fixed $P$ in the continuum of probability distributions. We prefer the first approach since it avoids the complication of dealing with the small differences between the actual type at a given $n$ and its limit $P$, which would require continuity arguments and dealing with undesired extra error terms. The same line of thought applies to the distortion measure $d$ in Theorem 1.

We choose $d$ to be a distortion measure with rational entries and consider the subsequence of $\{n\}$ for which $d$ belongs to a certain uniform grid in the space of distortion matrices, where the spacings between the grid points are proportional to $1/N_n$, where $N_n$ may tend to infinity even exponentially, but not faster than that. This will be useful as we will use the union bound within the grid and there will be no need to apply continuity arguments to pass to the continuum of matrices.

**III. The Probability of a Successful Single Random Selection**

This section is devoted to a lemma that stands at the heart of the derivations in this work: It provides an asymptotically tight assessment of the probability that a single randomly selected codeword happens to fall within distortion no more than $nD$ away from a given source vector $x \in \mathcal{X}^n$, which has a certain empirical distribution, $\hat{P}_\mathcal{X} = P$, $P \in P_n$. The concept of proving achievability of $R_d(D, P)$ via such a lower bound is, of course, by no means new, and it serves as the classical tool for proving the direct part of the rate-distortion coding theorem. There are two points, however, that make our derivation somewhat different from the traditional one.

1. We select a universal random coding distribution that is asymptotically as good as the optimal one for every source and every distortion measure.

2. Our analysis is based upon the saddle-point method (a.k.a. the steepest descent method) [10, Chap. 5], [19, Section 4.3], which is not only exponentially tight, but moreover, it is asymptotically tight in the sense that the ratio between the approximate expression and the exact probability tends to unity as $n \to \infty$. As a consequence, it gives rise to a precise characterization of the redundancy terms as well.

Consider the random coding distribution, given by the uniform mixture of all memoryless sources,

$$W(\hat{x}) = (K-1)! \cdot \int_{\mathcal{Q}} dQ \cdot \prod_{i=1}^{n} Q(\hat{x}_i), \quad (6)$$

where $Q$ is the simplex of all probability assignments over $\mathcal{X}$ and the factor $(K-1)!$ is a normalization constant that accounts for the fact the volume of $Q$ is $1/(K-1)!$. The probability of a successful single random selection, for a given source sequence $x$, is defined as

$$P^s_d[\hat{x}] = \sum_{\{\hat{x}: d(\hat{x}, x) \leq nD\}} W(\hat{x}) = (K-1)! \cdot \int_{\mathcal{Q}} dQ \cdot \sum_{\{\hat{x}: d(\hat{x}, x) \leq nD\}} \prod_{i=1}^{n} Q(\hat{x}_i).$$

Before stating our main lemma, we list certain assumptions and definitions, along with some background and a few observations.

1. For the case where the non-zero entries of the distortion matrix, $\{d(j, k), \ 1 \leq j \leq J, \ 1 \leq k \leq K\}$, are all commensurable, i.e., the ratios, $d(j, k)/d(j', k')$ ($d(j', k') \neq (j, k), \ d(j', k') > 0$) are all rational numbers, we define $\Delta$ as the greatest common factor of $\{d(j, k) : d(j, k) > 0, \ 1 \leq j \leq J, \ 1 \leq k \leq K\}$. In other words, $\Delta$ is the largest positive real, $\delta$, such that $d(j, k)/\delta$ is a positive integer for every $(j, k)$ with $d(j, k) > 0$. Otherwise, if the non-zero entries of the distortion matrix are incommensurable, we define $\Delta = 0$ (which amounts to passing to the limit $\Delta \to 0$).

2. For a given $P$ and $d$, let us define

$$D_{\max}(Q) \triangleq \sum_{x} \sum_{\hat{x}} P(x)Q(\hat{x})d(x, \hat{x}). \quad (7)$$

and consider the case where

$$D < D_{\max} \triangleq \min_{Q} D_{\max}(Q) = \min_{\hat{x}} \sum_{x} P(x)Q(\hat{x})d(x, \hat{x}) \quad (8)$$

2The choice of the uniform mixture is motivated merely by its convenience. It can be replaced by any density $w(Q)$, as long as it is bounded away from zero and from infinity.

3This well known fact can easily be proved either by induction on $K$ or by the simple observation that the volume occupied by the set of vectors, $(u_1, \ldots, u_{K-1})$, with ordered components, $0 \leq u_1 \leq u_2 \leq \ldots \leq u_{K-1} \leq 1$, which is obviously $1/(K-1)!$, can be transformed bijectively into a set of $K-1$ probabilities, $p_1 = u_1, p_2 \leq u_2 - u_1, \ldots, p_{K-1} = u_{K-1} - u_{K-2}$ (whose sum is $u_{K-1} \leq 1$), and that the Jacobian of this transformation is 1, so it does not alter the volume.
which implies \( R_d(D, P) > 0 \). Let \((s^*, Q^*)\) be the saddle-point of \( F(s, Q) \), that is,

\[
\begin{align*}
  s^* &= \arg \max_s \min_Q F(s, Q) \\
  Q^* &= \arg \min_s \max_Q F(s, Q).
\end{align*}
\]

If \( D < D_{\text{max}}(Q^*) \), then \( s = s^* \) is the solution to the equation

\[
\sum_x P(x) \cdot \frac{\sum_x Q^*(\hat{x}) e^{-s d(x, \hat{x})} d(x, \hat{x})}{\sum_x Q^*(\hat{x}) e^{-s d(x, \hat{x})}} = D,
\]

and it is easy to see that \( s^* > 0 \), due to the decreasing monotonicity of the l.h.s. of this equation as a function of \( s \) and the fact that for \( s = 0 \), it yields \( D_{\text{max}}(Q^*) \) which cannot be smaller than \( D_{\text{max}} \), by definition, and a-fortiori, it is strictly larger than \( D \).

3. Since the saddle-point method will be used, we need to consider the extension of the function \( F(s, Q) \) (as a function of \( s \)) to the complex plane, that is, the function \( F(z, Q) \) where \( z \) is a complex variable. We first need to establish the analyticity property of \( F(z, Q^*) \), at least in some neighborhood of the vertical line, \( z = s^* + jw, \omega \in \mathbb{R}, j \triangleq \sqrt{-1} \). The subtle issue here is that \( F \) includes the complex logarithm function, which is inherently discontinuous wherever its argument crosses the negative real axis in either direction. There is no single branch of the complex logarithm which is continuous (let alone, analytic) since the origin, which is a singular point of the logarithmic function, is surrounded, in general (possibly, infinitely many times) by the path drawn by \( A_z (\omega) \triangleq \sum_x Q^*(\hat{x}) e^{-(s^* + jw) d(x, \hat{x})} \) in the complex plane, as \( \omega \) varies along the real line.\(^4\) One way to alleviate the problem, is to proceed as follows. We take \( \arg A_z (0) = 0 \), corresponding to the principal branch of the complex logarithm, and for \( \omega > 0 \), we create a continuous (phase-unwrapped) version of \( \arg A_z (\omega) \) by removing all discontinuities upon adding or subtracting \( 2\pi \). In the language of complex analysis, this process amounts to tailoring neighboring branches of the complex logarithm, \( A_z (\omega) \), so as to preserve continuity and analyticity. Specifically, let \( \omega_1 \) denote the first positive value of \( \omega \) such that \( \arg A_z (\omega) \) either passes the level \( \pi \) upward or crosses the level \( -\pi \) downward. In the first case, this is formally defined as follows: for every \( \epsilon > 0 \) there exists \( \delta > 0 \), such that \( \omega_1 > \omega > \omega_1 - \epsilon \) implies \( \pi > A_z (\omega) > \pi - \delta \), and suppose that

\[
\frac{d \arg A_z (\omega)}{d \omega} \bigg|_{\omega = \pi} > 0
\]

(or if the first derivative is zero, consider the second derivative, etc.). Then, at \( \omega = \omega_1 \), we pass to the next branch of the complex logarithm by adding \( 2\pi \), thereby removing the discontinuity. Likewise, if \( A_z (\omega) \) crosses the level \( -\pi \) downward (with an analogous formal definition), we remove the discontinuity by subtracting \( 2\pi \). We then continue to increase \( \omega \) and follow the corresponding branch until the next point, \( \omega_2 \), at which \( \arg A_z (\omega) \) either crosses the level \( \pi \) upward or crosses \( -\pi \) downward, and so on. For \( \omega < 0 \), the process

\(^4\)This issue is intimately related to the well-known phase unwrapping problem (see, e.g., [1] and references therein).

is similar and can be obtained by the negative reflection of the unwrapped phase at \( \omega > 0 \).

4. Since \( z = s^* \) is a point where \( dF(z, Q^*)/dz = 0 \), it is a saddle-point of \( |e^{-nF(z, Q^*)}| = e^{-n\text{Re}(F(z, Q^*))} \) (and hence also of \( \text{Re}(F(z, Q^*)) \), in the complex plane. Therefore, since \( z = s^* \) maximizes \( \text{Re}(F(s, Q^*)) \equiv F(s, Q^*) \) along the real axis, it minimizes \( \text{Re}(F(z, Q^*)) \) (at least locally) along the vertical line, \( \text{Re}(z) = s^* \), and so, \( \text{Re}(F(s^* + j\omega, Q^*)) \), as a function of \( (\omega, Q^*) \), is minimized at \((0, Q^*)\). We assume that \( P, d \) and \( D \) are such that the \( K \times K \) Hessian, \( S \), of \( F(s^* + j\omega, Q^*) \) (as a function of \( (\omega, Q^*) \)), formed by the second-order partial derivatives of \( F \) w.r.t. \( \omega, Q(1), \ldots, Q(K-1) \), computed at \((0, Q^*)\), is a positive definite matrix. Let \( V(P, d) \) denote the absolute value of the determinant of the Hessian of \( F \) at \((0, Q^*)\).

5. A necessary condition (which is relatively easy to check) for the positivity of \( S \) at \((0, Q^*)\), is the positivity of the submatrix, \( Z \), of the second-order derivatives of \( F \) only w.r.t. the first \( K-1 \) components of \( Q \) at that point. A straightforward calculation shows that this submatrix can be represented as \( \sum_{x, x} a_{x} v_{x} e_{x}^{T} \), where \( \{a_{x}\} \) are positive reals and for every \( x \in \mathcal{X}, v_{x} \) is a \((K-1)\)-dimensional column vector, whose \( x \)-th component \((1 \leq x \leq K-1)\) is given by \( e^{-s_{d_{x}} d(x, z)} - e^{-s_{d(x, z)}} \), where we have taken \( \mathcal{X} \) to be \( \{1, 2, \ldots, K\} \). Therefore, if \( J \geq K-1 \) and \( \{v_{x}, x \in \mathcal{X} \} \) span the entire \((K-1)\)-dimensional vector space, the Hessian submatrix is positive definite. To complete the submatrix to the full Hessian of \( F \) as a function of \( (\omega, Q^*) \), we need the corner entry corresponding to \( \omega \), that is, \( h = \partial^2 F(s^* + j\omega, Q^*)/\partial \omega^2 \) and the \((K-1)\)-dimensional column vector, \( g \), whose \( x \)-th component is given by \( \partial^2 F(s^* + j\omega, Q^*)/\partial \omega \partial Q^*(\hat{x}) \), thus, partitioning the full Hessian into a block structure,

\[
S = \begin{pmatrix} h & g^T \\ g & Z \end{pmatrix}.
\]

If \( Z \) is positive definite, then it has an inverse, \( Z^{-1} \), and then the determinant of \( S \) is given by the well-known relation,

\[
|S| = |Z| \cdot (h - g^T Z^{-1} g).
\]

Therefore, if in addition to the positivity of \( Z \), we also have that \( h > g^T Z^{-1} g \), then \( S \) is also positive definite.

6. Referring to items 1 and 4 above, in the commensurable case, the function \( \text{Re}(F(s^* + j\omega, Q^*)) \) is periodic in \( \omega \) with a period given by \( \Omega = 2\pi/\Delta \), and then there are infinitely many minima, all given by \((k\Omega, Q^*)\), \( k = 0, \pm 1, \pm 2, \ldots \). Obviously, the Hessian at all these points are the same as in \((s^*, Q^*)\).

7. Finally, for the commensurable case, we define the function

\[
T_n(P, d) \triangleq \frac{(K-1)! \cdot (2\pi)^{K/2-1} \times \Delta \exp\{-s^*[(nD) \mod \Delta]\}}{(1 - e^{-s^*\Delta})\sqrt{V(P, d)}},
\]

where \( a \mod b \triangleq a - b \cdot \lfloor a/b \rfloor \). For, the incommensurable case, \( T_n(P, d) \) is defined by taking the limit \( \Delta \to 0 \), i.e.,

\[
T_n(P, d) \triangleq \frac{(K-1)! \cdot (2\pi)^{K/2-1}}{s^* \sqrt{V(P, d)}}.
\]
We are now ready to state the following lemma.

**Lemma 1:** Let $P \in \bigcup_{k=1}^{\infty} \mathcal{P}_k$ be given and let $d$ be a given distortion measure that satisfies assumptions of Section II. Let $\mathcal{N}$ be a sequence of positive integers, $\{n\}$, such that $P \in \mathcal{P}_n$. Then, the following two statements hold for every $x \in \mathcal{T}_n(P)$, $n \in \mathcal{N}$:

1. For $P$ and $d$ such that $R_d(D,P) > 0$, assume that the Hessian, $S$, of $F(s^* + j\omega, Q)$ w.r.t. $(\omega, Q(1), \ldots, Q(K-1))$, computed at $(0, Q^*)$, is positive definite. Then,

$$P_s^d[x] \geq \frac{T_{n}(P,d)}{\exp(nR_d(D,P))} \cdot (1 - \epsilon_{P,d}(n)), \tag{16}$$

where $\epsilon_{P,d}(n)$ denotes a (not necessarily positive) sequence, which depends on $P$ and $d$, but tends to zero for fixed $P$ and $d$, provided that $n \to \infty$ along $\mathcal{N}$.

2. For $P$ and $d$ such that $R_d(D,P) = 0$, $P_s^d[x] = (K - 1)! \cdot \text{Vol}(Q : D_{\text{max}}(Q) \leq D) \cdot (1 - \epsilon_{P,d}(n))$, where $\epsilon_{P,d}(n)$ has the same properties as $\epsilon_{P,d}(n)$.

**Discussion:** A few comments are in order concerning this lemma.

1. First, two technical issues should be clarified. The first concerns the formulation of the lemma for a given rational $P$ and along all values of $n$ for which $P \in \mathcal{P}_n$. This is done in order to fix $P$ while $n$ grows without bound, which is necessary in order to argue that $\epsilon_{P,d}(n) \to 0$, as this statement holds for every fixed $P$ (and $d$). Otherwise, if $n$ exhausts all positive integers, $P$ does not exist as an empirical distribution for most values of $n$. The second issue concerns the term $\sum_{j=0}^{K\ln n} \beta_{P,d}(n)$.

We conclude from the lemma that in the interesting case where $R_d(D,P) > 0$,

$$P_s^d[x] \geq \exp\left\{ - \left[nR_d(D,P) + \frac{K \ln n}{2} + \beta_{P,d}(n)\right]\right\}, \tag{17}$$

where $\beta_{P,d}(n) = -\log[T_n(P,d)(1 - \epsilon_{P,d}(n))]$ is upper bounded by a constant, for all sufficiently large $n$, but both the constant and the sufficiently large $n$ depend on $P$ and $d$. In the less interesting case of $R_d(D,P) = 0$, $P_s^d[x]$ is essentially a positive constant.

2. The choice of the mixture distribution (6) as our random coding distribution is inspired by [14], as well as on the intimately related problem of guessing, [7], [20], but here our analysis is more refined for the quest of quantifying rate redundancies. For a rough insight on the rationale behind this choice, consider the following line of thought. Intuitively, $W(\hat{x})$ is exponentially equivalent to the normalized maximum-likelihood (NML) distribution, that is proportional to $\max_{Q \in \mathcal{Q}} Q(\hat{x})$, whose normalization factor, $\sum_{\hat{x}} \max_{Q \in \mathcal{Q}} Q(\hat{x})$ (a.k.a. the Shtravok sum), grows only polynomially with $n$ (as can easily be seen by the method of types). Consequently, the probability of any $x$ under the NML distribution (and hence also under $W$), is exponentially no smaller than $Q(\hat{x})$ for every product distribution $Q$, including the optimal one. As a result, the probability of a single success under $W$ is exponentially no worse than the one induced by every product distribution $Q$. Indeed, we could have chosen our random distribution to be the NML distribution, but the mixture distribution, $W$, lends itself more conveniently to analysis. In fact, Mahmoud and Wagner [17] employed the NML distribution, but in a different way than here.

3. Earlier results on universal lossy coding were derived under different regularity conditions than those that we impose here. For example, in [31, Theorem 2] it is assumed that the Hessian of $R_d(D,P)$ w.r.t. the source letter probabilities, $\{Q(x), x \in X\}$, is uniformly bounded by a finite constant over a class, $\mathcal{F}$, of source distributions $P$, and that absolute value of the expected derivative w.r.t. $D$ of the empirical rate-distortion function is finite. In [30], it is assumed that the matrix $\{e^{a \cdot s(x)}, x \in X, \hat{x} \in X\}$ (s < 0) is of column full rank, namely, the rank is equal to $K$ and that a certain Jacobian would have a non-zero determinant. However, it is difficult to compare those sets of regularity conditions to ours. The conditions in [31] are more demanding in some aspects, and less demand in others, but [31] is not about pointwise redundancy and it does not address distortion-universality. The conditions in [30] appear less demanding than ours, but it does not guarantee that the code would be $d$-semifaithful, and it is not a pointwise result either. It is easy to see, for example, that the binary symmetric source with the Hamming distortion measure satisfies our regularity conditions.

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

**Proof of Lemma 1:** We assume that all non-zero entries of the distortion matrix $d$ are commensurable. The non-commensurable will be treated afterwards, where it will be shown that it can also be obtained in the limit of $\Delta \to 0$, or equivalently, $\Omega \to \infty$. Our proof is based on the following identity regarding the unit step function, $U(x) \triangleq \mathbb{I}(x \geq 0)$, which manifests the fact that it can be represented as the inverse Laplace transform (Mellin’s inverse formula) of the complex function $1/z = \int_{0}^{\infty} e^{-z} dz$ (Re$\{z\} > 0$):

$$U(x) = \frac{1}{2\pi j} \lim_{A \to \infty} \int_{c-jA}^{c+jA} \frac{e^{\pm x}}{z} \cdot dz, \quad x \neq 0, \tag{18}$$

where $j \triangleq \sqrt{-1}$, $A > 0$, and $c$ is an arbitrary positive real. For $x = 0$, on the other hand, $U(0)$, as defined, is equal to 1, whereas the the integral on the r.h.s. yields $1/2$. Therefore, strictly speaking, we have:

$$U(x) \geq \frac{1}{2\pi j} \lim_{A \to \infty} \int_{c-jA}^{c+jA} \frac{e^{\pm x}}{z} \cdot dz, \quad x \in \mathbb{R}. \tag{19}$$

We then have the following chain of equalities/inequalities:

$$P_s^d[x] = (K - 1)! \sum_{\{\hat{x} : d(\hat{x}, x) \leq nD\}} \int_{\mathcal{Q}} Q(\hat{x}) dQ$$

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Consider now the $\ell$-th term of this infinite summation, i.e.,

$$B_\ell = \frac{(K-1)!}{2\pi j} \int_{c+j(\ell-1/2)\Omega}^{c+j(\ell+1/2)\Omega} \frac{e^{-nF(z,Q)}}{z} \cdot dQdz,$$

(21)

which is a multi-dimensional integral of a complex function with one complex variable and $K-1$ real variables. As can be seen, the integration over the complex variable, $z$, is along the vertical line in the complex plane, from $c + j(\ell - 1/2)\Omega$ to $c + j(\ell + 1/2)\Omega$, where $c > 0$ is arbitrary. This integral will now be assessed using the saddle-point method (a.k.a. the steepest descent method), following the exposition in [5, Sections 4] derbruijn81. Consider the case where $D < D_{\text{max}}$, so that $s^* > 0$. In this case, we can select $c = s^*$, and then the integration path over $z$ would pass via the saddle-point $s^* + j\Omega$. The assumptions of the lemma allow to apply Theorem 2.1 of [23] (with the vector $(\omega, Q)$ playing the role of $t$ therein), in support of the application of the multivariate saddle-point theorem, to obtain:

$$B_\ell = \frac{(K-1)!}{2\pi j} \cdot j \cdot \frac{1}{s^* + j\Omega} \cdot \sqrt{\frac{(2\pi)^K}{nK^2 V(P,d)}} \cdot e^{-nF(s^*,\Omega^*)} \cdot \left[1 - e_{P,d}(n)\right]$$

$$= \frac{(K-1)!}{2\pi j} \cdot \frac{1}{s^* + j\Omega} \cdot \sqrt{\frac{(2\pi)^K}{nK^2 V(P,d)}} \cdot e^{-nF(s^*,\Omega^*)} \cdot \left[1 - e_{P,d}(n)\right]$$

$$= \frac{(K-1)!}{2\pi j} \cdot \frac{1}{s^* + j\Omega} \cdot \sqrt{\frac{(2\pi)^K}{nK^2 V(P,d)}} \cdot e^{-nF(s^*,\Omega^*)} \cdot \left[1 - e_{P,d}(n)\right],$$

(22)

where the second factor $j$ in the first line is due to the change in the integration of the complex variable according to $dz = j\omega$ (along the line $z = s^* + j\omega$). Since $P_{\ell}[x] \geq \sum_{\ell=-\infty}^{\infty} B_\ell$, it remains to calculate the sum of the factors that depend on $\ell$:

$$\sum_{\ell=-\infty}^{\infty} e^{j\Omega \ell \Omega n D} \cdot \frac{1}{s^* + j\Omega} \cdot \sum_{\ell=-\infty}^{\infty} \delta(\omega - \Omega t) \cdot d\omega$$

$$= 2\pi \cdot \left\{ \left[ e^{-s^* t U(t)} \right] \cdot \left[ \frac{1}{\Omega} \sum_{k=-\infty}^{\infty} \delta \left( t - \frac{2\pi k}{\Omega} \right) \right] \right\} \bigg|_{t=nD}$$

$$= 2\pi \cdot \sum_{k=-\infty}^{\infty} e^{-s^* (nD - 2\pi k/\Omega)} U \left( nD - 2\pi k/\Omega \right)$$

$$= 2\pi \cdot \exp \left\{ -s^* \left[ (nD) \mod (2\pi/\Omega) \right] \right\} \cdot \sum_{k=0}^{\infty} e^{-s^* \cdot 2\pi k/\Omega}$$

$$= 2\pi \cdot \exp \left\{ -s^* \left[ (nD) \mod (2\pi/\Omega) \right] \right\} \cdot \left[ \frac{\Omega}{1 - e^{-s^* \cdot 2\pi/\Omega}} \right]$$

$$= \frac{\Delta \exp \left\{ -s^* \left[ (nD) \mod (2\pi/\Omega) \right] \right\}}{1 - e^{-s^* \Delta}},$$

(23)

where in (a) we have used the fact that inverse Fourier transform of the product of two frequency-domain functions is equal to the convolution between the individual inverse Fourier transforms. If the positive distortions, $\{d(j,k)\}$, are incommensurable, then $\text{Re} (F(s^* + j\omega, Q))$ is no longer periodic and then only $z = s^*$ is a dominant saddle-point. In this case, instead of the infinite summation of (23), we have only the term corresponding to $\ell = 0$, which is $1/s^*$. This can be viewed as a special case of the right-most side of (23), pertaining to the limit $\Delta \to 0$ (or, equivalently, $\Omega \to \infty$), which matches the above formal definitions of $\Delta$ and $\Omega$ in...
the incommensurable case. Finally, we obtain
\[
\Pr\{I_d(x) > M\} = (1 - P^d[x])^M = \exp\{M \ln(1 - P^d[x])\} 
\leq \exp\{-M \cdot P^d[x]\},
\]
and so, if \(M = M_n = e^\lambda_n / P^d[x]\), for some arbitrary positive sequence, \(\lambda_n\), that tends to infinity, then
\[
\Pr\{I_d(x) > M_n\} \leq \exp(-e^\lambda_n).
\]
In particular, eq. (27) holds if \(R_d(D, P) > 0\) and \(M_n = \exp(nR_d(D, P) + K \ln n + \beta_p, d(n) + \lambda_n)\), or if \(R_d(D, P) = 0\) and \(M_n = \exp(\lambda_n + C)\), where \(C > 0\) is some constant. We will make use of this fact several times in this section.

For later use, we also need the following uniform lower bound to \(P^d[x]\): For a given \(x\), let \(\hat{x}_0 \in X^n\) denote the reproduction vector with the property \(d(\hat{x}, \hat{x}_0) = 0\), which exists by assumptions in Section II. Then,
\[
P^d[x] \geq (K - 1)! \int_{Q} \sum_{i \leq nD} Q(\hat{x})dQ 
\geq (K - 1)! \int_{Q} Q(\hat{x}_0)dQ 
\geq (K - 1)! \cdot \prod_{i \leq nD} [n! \hat{P}_d(\hat{x})]! / (n + K - 1)!
\]
\[
\geq \exp\left\{-nH(\hat{P}_{\hat{x}_0}) + \left(\frac{K - 1}{2}\right) \ln n + C\right\}
\geq \exp\left\{-n \ln K + \left(\frac{K - 1}{2}\right) \ln n + C\right\},
\]
where (a) follows from [15, eq. (2.3)], (b) follows from [15, eq. (2.7)], and \(C\) is a constant.

Consider a randomly selected codebook of \(A^n\) codewords, where \(A\) is an arbitrary positive integer, strictly larger than \(K\), and where each codeword is drawn independently under \(W\). Let the randomly selected codebook be revealed to both the encoder and the decoder.

Consider next the following encoder. Similarly as before, let \(I_d(x)\) be defined as the index of the first codeword that falls within \(d\)-distortion \(nD\) away from \(x\), but now, with the small twist that if none of the \(A^n\) codewords falls within distortion \(nD\) from \(x\), then we define \(I_d(x) = A^n\) nevertheless (even though the distortion is larger than \(nD\)). Define the following probability distribution over the integers, \(1, 2, \ldots, A^n\):
\[
u[i] = \frac{1}{\sum_{k=1}^{A^n} 1/k}, \quad i = 1, 2, \ldots, A^n.
\]
Given \(x\) and distortion measure \(d\), the encoder finds \(I_d(x)\) and encodes it using a variable-rate lossless code with the length function (in nats, and ignoring the equivalent of the integer length constraint),
\[
L_d(x) = -\ln u[I_d(x)] 
\leq \ln I_d(x) + \ln(\sum_{k=1}^{A^n} 1/k) 
\leq \ln I_d(x) + \ln(\ln A^n + 1) 
\leq \ln I_d(x) + \ln(n \ln A + 1) 
\leq \ln I_d(x) + \ln n + c,
\]
where \(c = \ln(\ln A + 1)\). Therefore, the expected codeword length for \(x \in T_n(P)\), w.r.t. the randomness of the code, is bounded by:
\[
E\{L_d(x)\} 
\leq E\{\ln I_d(x)\} + \ln n + c 
\leq \ln E\{I_d(x)\} + \ln n + c 
\leq \ln \left(\sum_{k=1}^{A^n} k \cdot \left(1 - P^d[x]\right)^{k-1} \cdot P^d[x]\right) + \ln n + c 
\leq \ln \left(\sum_{k=1}^{A^n} \min\{k, A^n\} \cdot \left(1 - P^d[x]\right)^{k-1} \cdot P^d[x]\right) + \ln n + c 
\leq \ln \left(\frac{1}{P^d[x]}\right) + \ln n + c
\]
\[
\leq nR_d(D, P) + \left(\frac{K}{2} + 1\right) \ln n +
\]
\[ \beta_{P,d}(n) + c, \]  
where in the last step, we have used eq. (17). For later use, we denote
\[ L_d^+(x) = \ln \left( \frac{1}{P_d^+(|x|)} \right) + \ln n + c, \]  
and so, we have shown that
\[ E\{L_d(x)\} \leq L_d^+(x). \]  
It should be kept in mind that \( L_d^+(x) \) depends on \( x \) only via its type.

Our goal, in this section, however, is more ambitious than achieving the result of (31). To describe it, we first define a fine grid, \( \mathcal{D}_n \), of distortion matrices, \( \{0,d_{\max}/N_n,2d_{\max}/N_n, \ldots, N_n d_{\max}/N_n\}_{d \times K} \) (or any other fine grid, or union of grids, with the same number of points) that comply with assumptions of Section II, where \( \{N_n\} \) may be any sequence of integers growing with \( n \) as \( o(\exp(n^{1+\epsilon})) \) for every \( \epsilon > 0 \).\(^5\) We wish to prove the existence of a codebook with the following properties: (a) \( L_d(x) \) is upper bounded in terms of \( R_d(D, \mathcal{P}_x) \) plus some redundancy terms for every \( x \in \mathcal{T}_n(P) \), every \( P \in \mathcal{P}_n \) and every \( d \in \mathcal{D}_n \); (b) The distortion constraint is met for every \( x \in \mathcal{X}^n \) and every \( d \in \mathcal{D}_n \). To prove the second property, our approach is similar to that of Mahmood and Wagner [16]: We consider the earlier defined fine grid, \( \mathcal{D}_n \), in the space of distortion matrices. In spite of the similarity to Mahmood and Wagner’s approach, there is an important difference: In our case, the quantization of the distortion measure takes part only in the proof itself, not in the actual codebook construction, as in [16].

Our main coding theorem, in this work, is the following.

**Theorem 1:** Let the assumptions of Lemma 1 hold and let \( \epsilon \) be an arbitrarily small positive real. Then, there exists a sequence of codebooks, \( \{C_n\} \), \( C_n = \{x_1, x_2, \ldots, x_{A^n}\} \subset \mathcal{X}^n \) and a sequence decoders, \( \{\psi_n\} \), such that for every \( d \in \bigcup_{k \geq 1} \mathcal{D}_k \), there exists a sequence of encoders, \( \{\phi_n\} \), such that for every \( P \in \bigcup_{k \geq 1} \mathcal{P}_k \), every \( n \in \mathcal{N} \), \( n \leq \hat{n} : d \in \mathcal{D}_n, \ P \in \mathcal{P}_n \) and \( x \in \mathcal{T}_n(P) \), the following two properties hold at the same time:

\( a) \) If \( R_d(D, P) > 0 \),
\[ L_d(x) \leq L_d^+(x) + (1 + \epsilon) \ln n \]
\[ \leq n R_d(D, P) + \left( \frac{K}{2} + 2 + \epsilon \right) \ln n + \beta_{P,d}(n) + c. \]

\( b) \) If \( R_d(D, P) = 0 \),
\[ L_d(x) \leq (1 + \epsilon) \ln n + \tilde{\beta}_{P,d}(n), \]
where
\[ \tilde{\beta}_{P,d}(n) = - \ln \left( (K - 1)! \cdot \text{Vol}\{Q : D_{\max}(Q) \leq D\} \times \right. \]
\[ \left. [1 - \epsilon P_d(n)] \right). \]

\(^5\) For example, \( N_n \) can even be an exponential function of \( n \) with an arbitrarily large (but fixed) exponential growth rate.

A few comments are in order before we turn to the proof of this theorem.

1. Note that the order of the universal and existential quantifiers in the theorem (i.e., the order of the quantifiers “for every” and “there exists”) is consistent with the universality properties described in Section II: The codebook and the decoder are doubly universal (in both \( P \) and \( d \)), whereas the encoding function is universal in \( P \), but not in \( d \).

2. The set \( \mathcal{N} \), defined in the theorem, includes at least all integer multiples of the least common multiple between \( n_1 = \min\{n : P \in \mathcal{P}_n\} \) and \( n_2 = \min\{n : d \in \mathcal{D}_n\} \).

3. The main redundancy term in part (a), namely,
\[ \left( \frac{K}{2} + 2 + \epsilon \right) \ln n/n, \]
should be compared with those of Mahmood and Wagner [16], where the coefficients in front of \( \ln n \) are, respectively, \( 2JK + J + 3 \), \( JK + J \), and \( J^2 K^2 + J - 2 \), in Theorems 1, 2, and 3 of [16]. The differences are quite significant, especially for large \( J \) and \( K \). We remind that these are sample-wise bounds, in contrast to the expectation results in [16] and [17], which are, however, uniform. On the other hand, our results are not minimax because of the last redundancy term that depends on the type of \( x \) and could be large for some particular sequences.

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

**Proof of Theorem 1:** In this proof, we confine attention only to the more interesting case where \( R_d(D, P) > 0 \), but the case \( R_d(D, P) = 0 \) can easily be handled in the very same manner. Consider the quantity
\[ E_n \]
\[ \Delta \mathbb{E} \left\{ \max \left( \max_{P \in \mathcal{P}_n} \max_{x \in \mathcal{T}_n(P)} \mathcal{I}(d(x, \hat{X}) > nD) \right) \right\}, \]
where the expectation is w.r.t. the randomness of the code, \( C_n \).

If we can bound \( E_n \) by a sequence, \( \delta_n \), that decays as \( n \to \infty \), this will imply that there must exist a code for which both
\[ \max_{P \in \mathcal{P}_n} \max_{x \in \mathcal{T}_n(P)} \mathcal{I}(d(x, \hat{x}) > nD) \leq \delta_n \]
and
\[ \max_{P \in \mathcal{P}_n} \max_{x \in \mathcal{T}_n(P)} (d(x, \hat{x}) - (1 + \epsilon) \ln n) \leq \delta_n \]
at the same time. Observe that since the left-hand side of (37) is either zero or one, then if we know that it must be less than \( \delta_n \to 0 \), for some codebook, \( C_n \), it means that it must vanish as soon as \( n \) is large enough such that \( \delta_n < 1 \), namely, \( d(x, \hat{x}) \leq nD \) for all \( x \) and \( d \in \mathcal{D}_n \). Also, by (38), for the same codebook, we must have
\[ L_d(x) \leq L_d^+(x) + (1 + \epsilon) \ln n + \delta_n, \]
for all \( P \in \mathcal{P}_n \), \( x \in \mathcal{T}_n(P) \), and \( d \in \mathcal{D}_n \), where the extra term, \( \delta_n \), adds a negligible amount to the redundancy.
To prove that $E_n$ decays, we begin with the simple fact that the maximum between two non-negative numbers is upper bounded by their sum, which implies that

$$E_n \leq E\left\{ \max_{P \in P_n} \max_{x \in T_n(P)} \max_{d \in D_n} I_d(x) > nD \right\} + E\left\{ \max_{P \in P_n} \max_{x \in T_n(P)} \max_{d \in D_n} \left( L_d(x) - L_d^+(x) - (1 + \epsilon) \ln n \right) \right\},$$

and so, it is enough to prove that each one of the terms decays with $n$. As for the first term, we have:

$$E\left\{ \max_{P \in P_n} \max_{x \in T_n(P)} \max_{d \in D_n} I_d(x, \hat{X}) > nD \right\} \leq E\left\{ \sum_{P \in P_n} \sum_{d \in D_n} \sum_{x \in T_n(P)} I_d(x, \hat{X}) > nD \right\} = \sum_{P \in P_n} \sum_{d \in D_n} \sum_{x \in T_n(P)} E\left\{ I_d(x, \hat{X}) > nD \right\} = \sum_{P \in P_n} \sum_{d \in D_n} \sum_{x \in T_n(P)} \Pr\{ d(x, \hat{X}) > nD \} \leq \sum_{P \in P_n} \sum_{d \in D_n} \sum_{x \in T_n(P)} \exp\left\{ -A^n P_d(x) \right\} \leq \sum_{P \in P_n} \sum_{d \in D_n} \sum_{x \in T_n(P)} \exp\left\{ -\exp\left\{ n \left[ \ln A - \ln K - \left( K - \frac{1}{2} \ln n - C_n \right) \right] \right\} \right\} \leq (n + 1)^{-1} \cdot \exp\left\{ -\exp\left\{ n \left[ \ln A - \ln K - \left( K - \frac{1}{2} \ln n - C_n \right) \right] \right\} \right\},$$

where in (a) we have used (28). This quantity decays double-exponentially rapidly as $n \to \infty$ since we have assumed $A > K$ and $N_n = o(\exp\{n^{1+\epsilon}\})$.

As for the second term of (40), we have:

$$E\left\{ \max_{P \in P_n} \max_{x \in T_n(P)} \max_{d \in D_n} \left( L_d(x) - L_d^+(x) - (1 + \epsilon) \ln n \right) \right\} = \int_0^\infty \Pr\{ \max_{P \in P_n} \max_{x \in T_n(P)} \max_{d \in D_n} \left( \ln I_d(x) - \ln \frac{1}{P_d(x)} - (1 + \epsilon) \ln n \right) \geq s \} ds = \int_0^{n \ln A} \Pr\{ \max_{P \in P_n} \max_{x \in T_n(P)} \max_{d \in D_n} \left( \ln I_d(x) - \ln \frac{1}{P_d(x)} \right) \geq c - c' \ln \ln n \} ds$$

where $c$ and $c'$ are constants that depends on $P$.

**Proof of Theorem 2:** Let $P \in P_n$ be given. Any $d$-semifathful code must fully cover the type class $T_n(P)$ with balls of radius $nD$ (henceforth, referred to as $D$-balls), centered at the various codewords. Let $\hat{x}_1, \ldots, \hat{x}_M \in \hat{X}^n$ be $M$ codewords. The number of members of $T_n(P)$ that are
covered by $\hat{x}_1, \ldots, \hat{x}_M \in \hat{X}^n$ is upper bounded as follows.

$$N(P, D) = \left| \bigcup_{i=1}^{M} \left[ T_n(P) \cap \{x : d(x, \hat{x}_i) \leq nD \} \right] \right|$$

$$\leq \sum_{i=1}^{M} \left| T_n(P) \cap \{x : d(x, \hat{x}_i) \leq nD \} \right|$$

$$\leq M \cdot \max_{\hat{x} \in \hat{X}^n} \left| T_n(P) \cap \{x : d(x, \hat{x}) \leq nD \} \right|$$

and so, the necessary condition for a complete cover, which is $N(P, D) \geq |T_n(P)|$, amounts to

$$M \geq \frac{|T_n(P)|}{\max_{\hat{x} \in \hat{X}^n} \left| T_n(P) \cap \{x : d(x, \hat{x}) \leq nD \} \right|} \triangleq M_0.$$  \hfill (43)

Consider now a variable-length code with a codebook of size $M$. Let $L(\hat{x}_i)$ denote the length (in nats) of the compressed binary string that represents $\hat{x}_i$. The number of codewords, $\{\hat{x}_i\}$, with $L(\hat{x}_i) \leq \ln M - \epsilon \ln n$ is upper bounded as follows:

$$|\{\hat{x} \in C_n : L(\hat{x}) \leq \ln M - \epsilon \ln n\}|$$

$$= \sum_{\ell=1}^{\ln M - \epsilon \ln n} |\{\hat{x} : L(\hat{x}) = \ell\}|$$

$$\leq \sum_{\ell=1}^{\ln M - \epsilon \ln n} e^\ell$$

$$= e^{\ln M - \epsilon \ln n}$$

$$< e^\epsilon$$

$$= e \cdot n^{-\epsilon} M,$$ \hfill (44)

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 $M(1 - e^{-\epsilon n^{-\epsilon}})$ out of the $M$ codewords in $C^n$ (that is, the vast majority codewords), we have

$$L(\hat{x}) \geq \ln M - \epsilon \ln n$$

$$\geq \ln M_0 - \epsilon \ln n$$

$$= \ln |T_n(P)| -$$

$$\max_{\hat{x} \in \hat{X}^n} \left| T_n(P) \cap \{x : d(x, \hat{x}) \leq nD \} \right| - \epsilon \ln n.$$  \hfill (45)

Now, it is well known and easy to show using the Stirling approximation (see also [6, p. 26, Problem 2.2]), that

$$\ln |T_n(P)| \geq \mathbb{E} \left[ d(X, \hat{X}) \right] \leq D$$

$$\leq \frac{1}{2} \ln n + c(P),$$ \hfill (46)

where $c(P)$ is a certain constant that depends only on $P$. To complete the proof, we invoke Lemma 3 of [32], which asserts that the logarithm of the size of the intersection between $T_n(P)$ and a $D$-ball is upper bounded by

$$\max_{\hat{x} \in \hat{X}^n} \left| T_n(P) \cap \{x : d(x, \hat{x}) \leq nD \} \right| \leq \max_{\hat{x} \in \hat{X}^n} \left| T_n(P) \cap \{x : d(x, \hat{x}) \leq nD \} \right| - \epsilon \ln n.$$  \hfill (47)

VI. OUTLOOK AND FUTURE DIRECTIONS - BEYOND MEMORYLESS SOURCES

Our results in Sections III and IV hold pointwise, for each and every individual source vector $x$, without taking the expectation w.r.t. the randomness of the source vector. But in spite of the pointwise nature of our results so far, the codes that we have been considering are still relevant only for the class of memoryless sources and additive distortion measures, since the length function, $L_d(x)$, whose main term is $nR_d(D, P)$, depends on $x$ only via its zeroth order empirical distribution, which is blind to any empirical dependencies and repetitive patterns within the source sequence, $x$.

In future work, we would like to remain in the realm of individual sequences, but to expand the scope to codes that are suitable beyond the memoryless structure, i.e., codes that are designed to exploit the memory within the given source sequence to be compressed by adopting the deeper and wider individual-sequence approach of Ziv and Lempel [33], and in the spirit of the parallel analysis in [20]. By the same token, we will be interested in more general classes of distortion measures, not necessarily additive ones.

ACKNOWLEDGMENT

The author wish to express his gratitude to the anonymous referees and the Associate Editor for the precious time they have devoted to read the manuscript carefully and for providing very useful comments.

REFERENCES

[1] H. Al-Nashi, “Phase unwrapping of digital signals,” IEEE Trans. Acoust., Speech Signal Process., vol. 37, no. 11, pp. 1693–1702, Nov. 1989.

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

[3] A. Barron, J. Rissanen, and B. Yu, “The minimum description length principle in coding and modeling,” IEEE Trans. Inf. Theory, vol. 44, no. 6, pp. 2743–2760, Oct. 1998.
[4] T. Berger, Rate Distortion Theory—A Mathematical Basis for Data Compression. Upper Saddle River, NJ, USA: Prentice-Hall, 1971.
[5] T. M. Cover and J. A. Thomas, Elements of Information Theory. Hoboken, NJ, USA: Wiley, 2006.
[6] I. Csiszár and J. Körner, Information Theory—Coding Theorems for Discrete Memoryless Systems. 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[7] A. Cohen and N. Merhav, “Universal randomized guessing subjected to distortion,” IEEE Trans. Inf. Theory, vol. 68, no. 12, pp. 7714–7734, Dec. 2022.
[8] R. G. Gallager, Information Theory and Reliable Communication. Hoboken, NJ, USA: Wiley, 1968.
[9] L. D. Davisson, “Universal noiseless coding,” IEEE Trans. Inf. Theory, vol. IT-29, no. 6, pp. 783–793, Nov. 1973.
[10] N. G. de Bruijn, Asymptotic Methods in Analysis. New York, NY, USA: Dover, 1981.
[11] R. G. Gallager, “Source coding with side information and universal coding,” Lab. Inf. Decis. Syst., MIT, Cambridge, MA, USA, Tech. Rep. LIDS-P-937, 1976.
[12] R. M. Gray, Source Coding Theory. Boston, MA, USA: Kluwer Academic, 1990.
[13] I. Kontoyiannis, “Pointwise redundancy in lossy data compression and universal lossy data compression,” IEEE Trans. Inf. Theory, vol. 46, no. 1, pp. 136–152, Jan. 2000.
[14] I. Kontoyiannis and J. Zhang, “Arbitrary source models and Bayesian codebooks in rate-distortion theory,” IEEE Trans. Inf. Theory, vol. 48, no. 8, pp. 2276–2290, Aug. 2002.
[15] R. Krichevsky and V. Trotimov, “The performance of universal encoding,” IEEE Trans. Inf. Theory, vol. IT-27, no. 2, pp. 199–207, Mar. 1981.
[16] A. Mahmod and A. B. Wagner, “Lossy compression with universal distortion,” 2021, arXiv:2110.07022.
[17] A. Mahmod and A. B. Wagner, “Minimax rate-distortion,” 2022, arXiv:2202.04481.
[18] N. Merhav, “A comment on ‘a rate of convergence result for a universal D-semifaithful code,’” IEEE Trans. Inf. Theory, vol. 41, no. 4, pp. 1200–1202, Jul. 1995.
[19] N. Merhav, “Statistical physics and information theory,” Found. Trends Commun. Inf. Theory, vol. 6, nos. 1–2, pp. 1–212, 2009.
[20] N. Merhav and A. Cohen, “Universal randomized guessing with application to asynchronous decentralized brute-force attacks,” IEEE Trans. Inf. Theory, vol. 66, no. 1, pp. 114–129, Jan. 2020.
[21] N. Merhav and M. Feder, “Universal prediction,” IEEE Trans. Inf. Theory, vol. 44, no. 6, pp. 2124–2147, Oct. 1998.
[22] N. Merhav and M. Feder, “A strong version of the redundancy-capacity theorem of universal coding,” IEEE Trans. Inf. Theory, vol. 41, no. 3, pp. 714–722, May 1995.
[23] T. Neuschel, “Apéry polynomials and the multivariate saddle point method,” Constructive Approximation, vol. 40, no. 3, pp. 487–507, Dec. 2014.
[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 Trans. Inf. Theory, vol. IT-30, no. 4, pp. 629–636, Jul. 1984.
[26] J. F. Silva and P. Plantanida, “On universal D-semifaithful coding for memoryless sources with infinite alphabets,” IEEE Trans. Inf. Theory, vol. 68, no. 4, pp. 2782–2800, Apr. 2022.
[27] V. N. Vapnik, Statistical Learning Theory. Hoboken, NJ, USA: Wiley, 1998.
[28] A. J. Viterbi and J. K. Omura, Principles of Digital Communication and Coding. New York, NY, USA: McGraw-Hill, 1979.
[29] M. J. Weinberger, N. Merhav, and M. Feder, “Optimal sequential probability assignment for individual sequences,” IEEE Trans. Inf. Theory, vol. 40, no. 2, pp. 384–396, Mar. 1994.
[30] E.-H. Yang and Z. Zhang, “The redundancy of source coding with a fidelity criterion—Part II: Coding at a fixed rate level with unknown statistics,” IEEE Trans. Inf. Theory, vol. 47, no. 1, pp. 126–145, Jan. 2001.
[31] B. Yu and T. P. Speed, “A rate of convergence result for a universal D-semifaithful code,” IEEE Trans. Inf. Theory, vol. 39, no. 3, pp. 813–820, May 1993.
[32] Z. Zhang, E.-H. Yang, and V. K. Wei, “The redundancy of source coding with a fidelity criterion. I. Known statistics,” IEEE Trans. Inf. Theory, vol. 43, no. 1, pp. 71–91, Jan. 1997.
[33] J. Ziv and A. Lempel, “Compression of individual sequences via variable-rate coding,” IEEE Trans. Inf. Theory, vol. IT-24, no. 5, pp. 530–536, Sep. 1978.

Neri Merhav (Life Fellow, IEEE) was born in Haifa, Israel, in March 1957. He received the B.Sc., M.Sc., and D.Sc. degrees in electrical engineering from the Technion—Israel Institute of Technology in 1982, 1985, and 1988, respectively.

From 1988 to 1990, he was with AT&T Bell Laboratories, Murray Hill, NJ, USA. Since 1990, he has been with the Electrical Engineering Department, Technion, where he is currently an Irving Shepard Professor. From 1994 to 2000, he was a Consultant with the Hewlett-Packard Laboratories—Israel (HPL-I). His research interests include information theory, statistical communications, statistical signal processing, the areas of lossless/lossy source coding and prediction/filtering, relationships between information theory and statistics, detection, estimation, and the area of Shannon theory, including topics in joint source-channel coding, source/channel simulation, coding with side information with applications to information hiding and watermarking systems, the relationships between information theory, and statistical physics.

Dr. Merhav was a co-recipient of the 1993 Paper Award of the IEEE Information Theory Society, the 1994 American Technion Society Award for Academic Excellence, and the 2002 Technion Henry Taub Prize for Excellence in Research. He also served as the Co-Chairperson of the Program Committee of the 2001 IEEE International Symposium on Information Theory. From 1996 to 1999, he served as an Associate Editor for the Special Issues on Source Coding of the IEEE TRANSACTIONS ON INFORMATION THEORY and for the Special Issues on Shannon Theory of the IEEE TRANSACTIONS ON INFORMATION THEORY from 2017 to 2020. He is currently on the Editorial Board of Foundations and Trends in Communications and Information Theory.