D-Semifaithful Codes that are Universal over Both Memoryless Sources and Distortion Measures

Neri Merhav

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

Abstract

We 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 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 distortion measure. As outlined in the last section, the principal ideas can also be extended significantly beyond the class of memoryless sources, namely, to the setting of individual sequences encoded by finite-state machines.

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

1 Introduction

We consider the classical problem of lossy compression for finite-alphabet memoryless sources with respect to a fidelity criterion defined by an additive distortion measure [2], [3, Chap. 10], [4, Chap. 9], [7], [18, Chaps. 7,8]. 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 [2], 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 rate-distortion function asymptotically, for long blocks, see, e.g., [8], [9], [10], [11], [15], [16], [19], [20], which is by no means an exhaustive list of all relevant articles. This line of research, along with its various types of universality (weak universality, strong universality, expected vs. almost-sure convergence, etc.) complements and partially extends its lossless counterpart, yet it should be pointed out that the theory of universal lossless source coding is significantly more mature and well developed, along with ties to other problem areas, such as channel capacity theory and universal prediction theory (see, for example, [14]).

In a recent work coauthored with Cohen [5] (which is a further development over [1] and [13]), 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 [5] are harnessed and considerably refined to demonstrate the existence of d-semifaithful codes, which are not only universal with respect to (w.r.t.) the source statistics, but also universal w.r.t. the class of all bounded 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 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. Recently, Mahmood and Wagner have also provided very interesting results along the very same line [9], [10]. In [9], they proposed three 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 Vapnik-Chervonenkis (VC) dimension [17]. 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 [10], as its title suggests, the focus is more towards strong universality and minimax properties of universal codes. Accordingly, several coding theorems are provided in [10], but the uniformity comes at the inevitable price of a slowdown in the decay of the rate redundancies.

Our approach is conceptually much simpler than those of [9] and [10], and we show that smaller rate redundancies are achievable. Moreover, the analysis is also simpler, as its main part is based on a saddle-point derivation of the probability that a randomly selected codeword would fall within distortion nD away from a source sequence of a given type class. This 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. However, for the sake of fairness, it must be pointed out that in contrast to [10], we make no claims concerning uniformity of convergence. Finally, we provide an informal outline of an extension of the main ideas beyond the realm of memoryless sources and additive distortion measures, as we consider individual source sequences encoded by finite-state machines, in the spirit of the Lempel-Ziv setting [22].

The outline of the remaining part of this paper is as follows. In Section 2, we establish the notation and formalize the problem. In Section 3, 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 4, we state and prove the main coding theorem concerning the universality. Finally, in Section 5, we consider the broader setup mentioned above.

2 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 $\boldsymbol{X} = (X_1, \ldots, X_n)$, ($n$ – positive integer) may take a specific vector value $\boldsymbol{x} = (x_1, \ldots, x_n)$ 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 channels will be denoted by the letter P or Q. 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 notation $a_n \sim b_n$, for two positive sequences, will stand for the property that $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$. 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. The empirical distribution of a sequence $\boldsymbol{x} \in \mathcal{X}^n$, which will be denoted by $\hat{P}_{\boldsymbol{x}}$, is the vector of relative frequencies $\hat{P}_{\boldsymbol{x}}(x)$ of each symbol $x \in \mathcal{X}$ in $\boldsymbol{x}$.

Let $X_1, X_2, \ldots$ be independent copies of a random variable (RV) X, taking on values in a finite alphabet $\mathcal{X} = \{1, 2, \ldots, K\}$, where $K > 1$ is a positive integer. We denote the distribution of X by $P = \{P(x),\ x \in \mathcal{X}\}$, where $P(x) \stackrel{\triangle}{=} \Pr\{X = x\}$. Let $\hat{\mathcal{X}} = \{1, 2, \ldots, J\}$ denote a finite reconstruction alphabet, where $J > 1$ is also a positive integer. A distortion measure $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ is a non-negative function of pairs $(x, \hat{x}) \in \mathcal{X} \times \hat{\mathcal{X}}$, which can also be thought of as a $K \times J$ matrix whose (j, k)-th entry is given by $d(j, k)$, $1 \leq j \leq J$, $1 \leq k \leq K$. We assume that the distortion measure d satisfies two requirements:

(i) For every $x \in \mathcal{X}$, $\min_{\hat{x}} d(x, \hat{x}) = 0$;

(ii) $d_{\max} \stackrel{\triangle}{=} \max_{(x,\hat{x})\mathcal{X}\times\hat{\mathcal{X}}} d(x, \hat{x}) < \infty$.

Note that (i) does not limit the generality, as every distortion measure can be modified so as to satisfy (i) without changing the essence. This is done by defining $d'(x, \hat{x}) = d(x, \hat{x}) - \min_{\hat{x}} d(x, \hat{x})$, shifting the distortion level D to $D' = D - \boldsymbol{E}\{\min_{\hat{x}} d(X, \hat{x})\}$, and observing that the shift, $\boldsymbol{E}\{\min_{\hat{x}} d(X, \hat{x})\}$, depends only on the source P and the distortion measure, not on the code.

The distortion between two vectors $\boldsymbol{x} \in \mathcal{X}^n$ and $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}^n$ will be defined additively as

$$d(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \sum_{i=1}^{n} d(x_i, \hat{x}_i). \tag{1}$$

A block code of length n consists of an encoder and a decoder. We consider a variable-rate encoder, which is a mapping, $\phi_n : \mathcal{X}^n \to \{0,1\}^*$, that maps the space of source vectors of length n, $\mathcal{X}^n$, into a set, $\{0,1\}^*$, of variable-length compressed bit strings. The decoder is a mapping, $\psi_n : \{0,1\}^* \to \mathcal{C}_n \subseteq \hat{\mathcal{X}}^n$, that maps the space of compressed strings into a codebook, $\mathcal{C}_n$, which is a certain subset of the reproduction space, $\hat{\mathcal{X}}^n$. The length (in nats) of $\phi_n(\boldsymbol{x})$ will be denoted by $L_d(\boldsymbol{x})$, where the subscript d denotes the distortion measure.[1] The coding rate for $\boldsymbol{x}$ is $L_d(\boldsymbol{x})/n$.

A code is called d-semifaithful w.r.t. a given distortion level D, if for every $\boldsymbol{x} \in \mathcal{X}^n$,

$$d(\boldsymbol{x}, \psi_n(\phi_n(\boldsymbol{x}))) \le nD. \tag{2}$$

As is well known, the rate-distortion coding theorem asserts that for a given memoryless source P and distortion measure d, there exist d-semifaithful codes, (ϕ_n, ψ_n), w.r.t. distortion level D, whose average coding rate R is arbitrarily close to

$$R_d(D, P) \stackrel{\Delta}{=} \min_{\{P_{\hat{X}|X}:\ \boldsymbol{E}\{d(X,\hat{X}) \le D\}} I(X; \hat{X}), \tag{3}$$

for all sufficiently large n. On the other hand, the converse theorem asserts that there are no d-semifaithful 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., [7, p. 90, Corollary 4.2.3]):

$$\begin{aligned} R_d(D, P) &= \sup_{s \ge 0} \min_{Q} \left\{ -\sum_{x \in \mathcal{X}} P(x) \ln \left[\sum_{\hat{x} \in \hat{\mathcal{X}}} Q(\hat{x}) e^{-sd(x,\hat{x})} \right] - sD \right\} \\ &= \min_{Q} \sup_{s \ge 0} \left\{ -\sum_{x \in \mathcal{X}} P(x) \ln \left[\sum_{\hat{x} \in \hat{\mathcal{X}}} Q(\hat{x}) e^{-sd(x,\hat{x})} \right] - sD \right\}, \end{aligned} \tag{4}$$

where minimization is over all probability assignments, $Q = \{Q(\hat{x}),\ \hat{x} \in \hat{\mathcal{X}}\}$, across the reproduction alphabet, $\hat{\mathcal{X}}$. Here, the second equality holds since the function,

$$F(s, Q) \stackrel{\Delta}{=} -\sum_{x \in \mathcal{X}} P(x) \ln \left[\sum_{\hat{x} \in \hat{\mathcal{X}}} Q(\hat{x}) e^{-sd(x,\hat{x})} \right] - sD \tag{5}$$

[1] The need for the subscript d will become clear in the sequel.

is convex in Q and concave in s.

Our objective is to prove that there exists a sequence of codes, $\{(\phi_n, \psi_n)\}_{n\geq 1}$ that are simultaneously d-semifaithful w.r.t. D for every distortion measure d that satisfies requirements (i) and (ii), and, at the same time, their code-length functions are arbitrarily close to $R_d(D, \hat{P}_{\boldsymbol{x}})$ for all $\boldsymbol{x} \in \mathcal{X}^n$ when n is sufficiently large n. We will also focus on the achievable redundancy as a function of n.

3 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 $\boldsymbol{x} \in \mathcal{X}^n$, which has a certain empirical distribution, $\hat{P}_{\boldsymbol{x}}$. The concept of proving achievability of $R_d(D, P)$ via the 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) [6, Chap. 5], [12, 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[2] mixture of all memoryless sources,

$$W(\hat{\boldsymbol{x}}) = (J-1)! \cdot \int_{\mathcal{Q}} \mathrm{d}Q \cdot \prod_{i=1}^{n} Q(\hat{x}_i), \tag{6}$$

[2]The 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.

where $\mathcal{Q}$ is the simplex of all probability assignments over $\hat{\mathcal{X}}$ and the factor $(J-1)!$ is a normalization constant that accounts for the fact the volume of $\mathcal{Q}$ is $1/(J-1)!$.[3] The probability of a *successful single random selection*, for a given source sequence $\boldsymbol{x}$, is defined as

$$P_{\mathrm{s}}^{d}[\boldsymbol{x}] \stackrel{\Delta}{=} \sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}} W(\hat{\boldsymbol{x}}) = (J-1)!\cdot\int_{\mathcal{Q}} \mathrm{d}Q \sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}} \prod_{i=1}^{n} Q(\hat{x}_i). \tag{7}$$

Before stating our main lemma, we need a few more definitions.

1. For the case where the non-zero entries of the distortion matrix, $\{d(j,k),\ 1\le j\le J,\ 1\le k\le K\}$, are all commensurable, i.e., the ratios, $d(j,k)/d(j',k')$ ($(j',k')\ne(j,k)$, $d(j',k')>0$) are all rational numbers, we define Δ as the greatest common factor of $\{d(j,k):\ d(j,k)>0,\ 1\le j\le J,\ 1\le k\le K\}$. In other words, Δ is the largest positive real, δ, 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 $Q\in\mathcal{Q}$, let s_0 be the (unique) maximizer of $F(s,Q)$ (defined in eq. (5)) in the range $s\ge 0$, which is given as follows. If $D<D_{\max}(Q)\stackrel{\Delta}{=}\sum_{x,\hat{x}}P(x)Q(\hat{x})d(x,\hat{x})$, then s_0 is the solution s to the equation

$$\sum_x P(x)\cdot\frac{\sum_{\hat{x}}Q(\hat{x})e^{-sd(x,\hat{x})}d(x,\hat{x})}{\sum_{\hat{x}}Q(\hat{x})e^{-sd(x,\hat{x})}} = D. \tag{8}$$

Note that s_0 depends on Q, and accordingly, in the sequel, we will denote it sometimes as $s_0(Q)$, especially in places where it will be important to emphasize this dependence. If $D\ge D_{\max}(Q)$, $s_0=0$. For $s>0$, we define $M(s,Q)$ as the absolute value of the second derivative of $F(s,Q)$ w.r.t. s. Let Q_0 be the minimizer of $F(s_0(Q),Q)$. For $D<D_{\max}(Q_0)$, we define $|\mathrm{Hess}_F(Q_0)|$ as the determinant of the $(J-1)\times(J-1)$ Hessian matrix of $F(s_0(Q),Q)$ w.r.t. the (first) $J-1$ components of Q, computed at $Q=Q_0$. Finally, define the function

$$K_n[s,Q]\stackrel{\Delta}{=}\frac{\Delta\exp\{-s[(nD)\bmod\Delta]\}}{(1-e^{-s\Delta})\sqrt{2\pi M(s,Q)}}, \tag{9}$$

where $a \bmod b \stackrel{\Delta}{=} a-b\cdot\lfloor a/b\rfloor$. We are now ready to state the following lemma.

[3] This well known fact can easily be proved either by induction on J or by the simple observation that the volume occupied by the set of vectors, $(u_1,\ldots,u_{J-1})$, with ordered components, $0\le u_1\le u_2\le\ldots\le u_{J-1}\le 1$, which is obviously $1/(J-1)!$, can be transformed bijectively into a set of $J-1$ probabilities, $p_1=u_1$, $p_2\le u_2-u_1$, ..., $p_{J-1}=u_{J-1}-u_{J-2}$ (whose sum is $u_{J-1}\le 1$), and that the Jacobian of this transformation is 1, so it does not alter the volume.

Lemma 1 *Let the assumptions of Section 2 hold. Then,*

$$P^d_{\mathrm{s}}[\boldsymbol{x}] \sim \begin{cases} (J-1)! \cdot \frac{(2\pi)^{(J-1)/2} K_n[s_0(Q_0), Q_0]}{\sqrt{|Hess_F(Q_0)|}} \cdot \frac{\exp\{-nR_d(D, \hat{P}_{\boldsymbol{x}})\}}{n^{J/2}} & R_d(D, \hat{P}_{\boldsymbol{x}}) > 0 \\ (J-1)! \cdot Vol\{Q: \ D_{\max}(Q) \le D\} \cdot [1 - o(n)] & R_d(D, \hat{P}_{\boldsymbol{x}}) = 0, \end{cases} \tag{10}$$

where $\hat{P}_{\boldsymbol{x}}$ denotes the empirical distribution of $\boldsymbol{x} \in \mathcal{X}^n$.

Discussion. A few comments are in order concerning Lemma 1.

1. First, a technical issue should be clarified. Note that although the factor $K_n[s_0(Q_0), Q_0]$ depends on n, it does not tend to zero as $n \to \infty$ and hence does not affect the asymptotic behavior for large n. Referring to eq. (9), this is easily seen by observing that the only dependence on n is in the exponential term of the numerator, which oscillates between $e^{-s\Delta}$ and 1. We therefore conclude that in the interesting case where $R_d(D, \hat{P}_{\boldsymbol{x}}) > 0$,

$$P^d_{\mathrm{s}}[\boldsymbol{x}] \sim \exp\left\{-n\left[R_d(D, \hat{P}_{\boldsymbol{x}}) + \frac{J \ln n}{2n} + o\left(\frac{\ln n}{n}\right)\right]\right\}. \tag{11}$$

For $R_d(D, \hat{P}_{\boldsymbol{x}}) = 0$, $P^d_{\mathrm{s}}[\boldsymbol{x}]$ is essentially a positive constant.

2. The choice of the mixture distribution (6) as our random coding distribution is inspired by earlier works on the intimately related problem of guessing, [5], [13], 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, $W(\boldsymbol{x})$ is exponentially equivalent to the normalized maximum-likelihood (NML) distribution, that is proportional to $\max_{Q \in \mathcal{Q}} Q(\hat{\boldsymbol{x}})$, whose normalization factor, $\sum_{\boldsymbol{x}} \max_{Q \in \mathcal{Q}} Q(\hat{\boldsymbol{x}})$ (a.k.a. the Shtarkov sum), grows only polynomially with n (as can easily be seen by the method of types). Consequently, the probability of any $\boldsymbol{x}$ under the NML distribution (and hence also under W), is exponentially no smaller than $Q(\hat{\boldsymbol{x}})$ 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, Mahmood and Wagner [10] employed the NML distribution, but in a different way than here.

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

Proof of Lemma 1. We begin with an evaluation of the probability of a single success under a given memoryless Q, leaving the integration over $\mathcal{Q}$ for the next step. Our proof is based on the following identity regarding the unit step function, $u(x) \stackrel{\Delta}{=} \mathcal{I}\{x \geq 0\}$, which manifests the fact that it can be represented as the inverse Laplace transform (Mellin's inverse formula) of the complex function $1/z = \int_0^\infty e^{-zx}\mathrm{d}z$ ($\mathrm{Re}\{z\} > 0$):

$$u(x) = \frac{1}{2\pi j}\lim_{A\to\infty}\int_{c-jA}^{c+jA}\frac{e^{zx}}{z}\cdot\mathrm{d}z, \tag{12}$$

where $j \stackrel{\Delta}{=} \sqrt{-1}$ and c is an arbitrary positive real. We then have the following chain of equalities:

$$\begin{aligned}
&\sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\leq nD\}} Q(\hat{\boldsymbol{x}}) \\
&= \sum_{\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}^n} Q(\hat{\boldsymbol{x}})\cdot u\left(nD - \sum_{i=1}^n d(x_i,\hat{x}_i)\right) \\
&= \sum_{\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}^n} Q(\hat{\boldsymbol{x}})\cdot\frac{1}{2\pi j}\lim_{A\to\infty}\int_{c-jA}^{c+jA}\frac{\mathrm{d}z}{z}\exp\left\{z\left(nD - \sum_{i=1}^n d(x_i,\hat{x}_i)\right)\right\} \\
&= \frac{1}{2\pi j}\lim_{A\to\infty}\int_{c-jA}^{c+jA}\frac{\mathrm{d}z}{z}e^{znD}\sum_{\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}^n} Q(\hat{\boldsymbol{x}})\cdot\exp\left\{-z\sum_{i=1}^n d(x_i,\hat{x}_i)\right\} \\
&= \frac{1}{2\pi j}\lim_{A\to\infty}\int_{c-jA}^{c+jA}\frac{\mathrm{d}z}{z}e^{znD}\prod_{x\in\mathcal{X}}\left[\sum_{\hat{x}\in\hat{\mathcal{X}}} Q(\hat{x})e^{-zd(x,\hat{x})}\right]^{n\hat{P}_{\boldsymbol{x}}(x)} \\
&= \frac{1}{2\pi j}\lim_{A\to\infty}\int_{c-jA}^{c+jA}\frac{\mathrm{d}z}{z}\exp\left\{n\left(zD + \sum_{x\in\mathcal{X}}\hat{P}_{\boldsymbol{x}}(x)\ln\left[\sum_{\hat{x}} Q(\hat{x})e^{-zd(x,\hat{x})}\right]\right)\right\} \\
&= \frac{1}{2\pi j}\lim_{A\to\infty}\int_{c-jA}^{c+jA}\frac{e^{-nF(z,Q)}}{z}\cdot\mathrm{d}z.
\end{aligned} \tag{13}$$

The right–most side of this chain of equalities is an integral of an exponential function with a large parameter n, along the vertical line in the complex plane, $\mathrm{Re}\{z\} = c$. This integral will now be assessed using the saddle-point method.

Consider the case where Q is such that $D < D_{\max}(Q)$, so that $s_0 > 0$. Suppose first that the positive entries of the distortion matrix are commensurable with a greatest common factor given by $\Delta > 0$. Since all non-zero $\{d(j,k)\}$ are integer multiples of Δ, the function $|e^{-nF(s_0+j\omega,Q)}| = \exp[-n\mathrm{Re}\{F(s_0+j\omega,Q)\}]$ is periodic in ω with period $\Omega \stackrel{\Delta}{=} 2\pi/\Delta$. Therefore, in the limit of $A \to \infty$, there are infinitely many dominant saddle-points, all of the form $z = s_0 + j\Omega\ell$, $\ell = 0, \pm 1, \pm 2, \ldots$, as in all these points, $|e^{-nF(z,Q)}|$ has a local maximum in the vertical direction of the complex plane (which is a global maximum within each period), and a minimum along the horizontal axis. In order that the integration path, $\text{Re}\{z\} = c$, would pass via all saddle-points, we select $c = s_0$. Thus, according to the saddle-point method [6, Chap. 5], [12, Sect. 4.3], in this case, we have

$$\begin{aligned} \sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}} Q(\hat{\boldsymbol{x}}) &\sim \frac{e^{j\pi/2}}{2\pi j}\sum_{\ell=-\infty}^{\infty}\frac{\exp\{-nF(s_0+j\Omega\ell,Q)\}}{s_0+j\Omega\ell}\cdot\sqrt{\frac{2\pi}{M(s_0+j\Omega\ell,Q)n}} \\ &= \left(\frac{1}{2\pi}\sum_{\ell=-\infty}^{\infty}\frac{e^{j\Omega\ell nD}}{s_0+j\Omega\ell}\right)\cdot\exp\{-nF(s_0,Q)\}\cdot\sqrt{\frac{2\pi}{M(s_0,Q)n}}, \end{aligned} \tag{14}$$

where in the asymptotic equality step, we have collected the contributions of all dominant saddle-points along the integration path from $s_0 - j\infty$ to $s_0 + j\infty$ (where the factor $e^{j\pi/2} = j$ accounts for the vertical axis of all saddle-points), and then, in the next equality, we have used the periodicity of $e^{-n\text{Re}\{F(z,Q)\}}$ (and hence also of its second derivative) in the vertical direction. We next address the infinite summation in the brackets of the last line of (14).

$$\begin{aligned} \frac{1}{2\pi}\sum_{\ell=-\infty}^{\infty}\frac{e^{j\Omega\ell nD}}{s_0+j\Omega\ell} &= \frac{1}{2\pi}\int_{-\infty}^{\infty}e^{j\omega nD}\cdot\frac{1}{s_0+j\omega}\cdot\left[\sum_{\ell=-\infty}^{\infty}\delta(\omega-\Omega\ell)\right]\mathrm{d}\omega \\ &\stackrel{\text{(a)}}{=} \left\{\left[e^{-s_0 t}u(t)\right]\star\left[\frac{1}{\Omega}\sum_{k=-\infty}^{\infty}\delta\left(t-\frac{2\pi k}{\Omega}\right)\right]\right\}\Bigg|_{t=nD} \\ &= \frac{1}{\Omega}\sum_{k=-\infty}^{\infty}e^{-s_0(nD-2\pi k/\Omega)}u\left(nD-\frac{2\pi k}{\Omega}\right) \\ &= \frac{1}{\Omega}\exp\left\{-s_0\left[(nD)\text{mod}\left(\frac{2\pi}{\Omega}\right)\right]\right\}\cdot\sum_{k=0}^{\infty}e^{-s_0 2\pi k/\Omega} \\ &= \frac{\exp\left\{-s_0\left[(nD)\text{mod}\left(\frac{2\pi}{\Omega}\right)\right]\right\}}{\Omega(1-e^{-2\pi s_0/\Omega})} \\ &= \frac{\Delta\exp\{-s_0[(nD)\text{ mod }\Delta]\}}{2\pi(1-e^{-s_0\Delta})}, \end{aligned} \tag{15}$$

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 $\text{Re}\{F(s_0+j\omega,Q)\}$ is no longer periodic and then only $z = s_0$ is a dominant saddle-point. This can be viewed as a special case pertaining to the limit $\Delta \to 0$ (or, equivalently, $\Omega \to \infty$), which matches the above formal definitions of Δ and Ω in the incommensurable case. On substituting the right-most side of (15) back into

(14), we obtain

$$\sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}} Q(\hat{\boldsymbol{x}}) \sim \frac{\Delta\exp\{-s_0[(nD) \bmod \Delta]\}}{2\pi(1-e^{-s_0\Delta})}\cdot\sqrt{\frac{2\pi}{M(s_0,Q)n}}\cdot\exp\{-nF(s_0,Q)\}$$
$$= K_n[s_0,Q]\cdot\frac{\exp\{-nF(s_0,Q)\}}{\sqrt{n}}. \tag{16}$$

In the case where Q is such that $D > D_{\max}(Q)$,

$$\sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}} Q(\hat{\boldsymbol{x}}) = 1 - \sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})> nD\}} Q(\hat{\boldsymbol{x}}) = 1-o(n), \tag{17}$$

by the weak law of large numbers. However, unless $R_d(D,\hat{P}_{\boldsymbol{x}}) = 0$, there is no Q for which $D > D_{\max}(Q)$.

We now move on to the second step, of integration over $\mathcal{Q}$, which will be carried out using the multivariate version of the Laplace method of integration (see, e.g., [6, Chap. 4], [12, Section 4.2]). Assuming that $R_d(D,\hat{P}_{\boldsymbol{x}}) > 0$,

$$\begin{aligned}
&\sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}} W(\hat{\boldsymbol{x}})\\
= \;& (J-1)!\cdot\int_{\mathcal{Q}}\sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}} Q(\hat{\boldsymbol{x}})\mathrm{d}Q\\
\sim \;& (J-1)!\cdot\int_{\mathcal{Q}}\mathrm{d}Q\cdot\frac{K_n[s_0(Q),Q]}{\sqrt{n}}\cdot e^{-nF(s_0(Q),Q)}\\
= \;& (J-1)!\cdot\left(\frac{2\pi}{n}\right)^{(J-1)/2}\cdot\frac{1}{\sqrt{|\mathrm{Hess}_F(Q_0)|}}\frac{K_n[s_0(Q_0),Q_0]}{\sqrt{n}}\cdot e^{-n\sup_{s\ge 0}F(s,Q_0)}\\
= \;& (J-1)!\cdot\left(\frac{2\pi}{n}\right)^{(J-1)/2}\cdot\frac{1}{\sqrt{|\mathrm{Hess}_F(Q_0)|}}\frac{K_n[s_0(Q_0)),Q_0]}{\sqrt{n}}\cdot e^{-n\min_Q\sup_{s\ge 0}F(s,Q)}\\
= \;& \frac{(J-1)!\cdot(2\pi)^{(J-1)/2}K_n[s_0(Q_0),Q_0]}{\sqrt{|\mathrm{Hess}_F(Q_0)|}}\cdot\frac{e^{-nR_d(D,\hat{P}_{\boldsymbol{x}})}}{n^{J/2}}
\end{aligned} \tag{18}$$

When $R_d(D,\hat{P}_{\boldsymbol{x}}) = 0$, we have

$$\begin{aligned}
P_{\mathrm{s}}^d[\boldsymbol{x}] &\sim \sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}} W(\hat{\boldsymbol{x}})\\
&\sim (J-1)!\cdot\int_{\{Q:\ D_{\max}(Q)<D\}}\mathrm{d}Q[1-o(n)]
\end{aligned}$$

$$= \quad (J-1)! \cdot \text{Vol}\{Q: \; D_{\max}(Q) < D\} \cdot [1 - o(n)]. \tag{19}$$

This completes the proof of Lemma 1.

4 Main Result

In the previous section, we focused on the evaluation of the probability that a single randomly chosen codeword, under the mixture distribution, happens to be successful in encoding a given source sequence, $\boldsymbol{x}$, within distortion nD. In this section, we harness the result of Lemma 1 for our main coding theorem. The analysis, in this section, will be based on the following simple well known fact: Let $\boldsymbol{x} \in \mathcal{X}^n$ be given and let $\hat{\boldsymbol{X}}_1, \hat{\boldsymbol{X}}_2, \ldots$ be a sequence of n-vectors in $\hat{\mathcal{X}}^n$, randomly and independently drawn under W. Let $I_d(\boldsymbol{x})$ denote the index, i, of the first vector $\hat{\boldsymbol{X}}_i$ with $d(\boldsymbol{x}, \hat{\boldsymbol{X}}_i) \leq nD$. Then, for every positive integer, M:

$$\Pr\{I_d(\boldsymbol{x}) > M\} = (1 - P_{\rm s}^d[\boldsymbol{x}])^M = \exp\{M \ln(1 - P_{\rm s}^d[\boldsymbol{x}])\} \leq \exp\{-M \cdot P_{\rm s}^d[\boldsymbol{x}])\}, \tag{20}$$

and so, if $M = M_n = e^{\lambda_n}/P_{\rm s}^d[\boldsymbol{x}]$, for some arbitrary positive sequence, $\{\lambda_n\}$, that tends to infinity, then

$$\Pr\{I_d(\boldsymbol{x}) > M_n\} \leq \exp\{-e^{\lambda_n}\}. \tag{21}$$

In particular, eq. (21) holds if $R_d(D, \hat{P}_{\boldsymbol{x}}) > 0$ and $M_n = \exp\{nR_d(D, \hat{P}_{\boldsymbol{x}}) + \frac{J \ln n}{2} + C + \lambda_n\}$, or if $R_d(D, \hat{P}_{\boldsymbol{x}}) = 0$ 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.

Consider next a randomly selected codebook of A^n codewords, where A in an arbitrary positive integer, strictly larger than $\max\{J, K\}$, 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 $I_d(\boldsymbol{x})$ be defined as the index of the first codeword that falls within d-distortion nD away from $\boldsymbol{x}$, but now, with the small twist that if none of the A^n codewords fall within distortion nD from $\boldsymbol{x}$, then we define $I_d(\boldsymbol{x}) = L^n$ nevertheless (even though the distortion is larger than nD). Define the following probability distribution over the integers, $1, 2, \ldots, A^n$:

$$U[i] = \frac{1/i}{\sum_{k=1}^{A^n} 1/k}, \qquad i = 1, 2, \ldots, A^n. \tag{22}$$

Given $\boldsymbol{x}$ and distortion measure d, the encoder finds $I_d(\boldsymbol{x})$ and encodes it using a variable-rate lossless code with the length function (in nats, and ignoring the equivalent of the integer length constraint),

$$\begin{aligned} L_d(\boldsymbol{x}) &= -\ln U[I_d(\boldsymbol{x})] \\ &\leq \ln I_d(\boldsymbol{x}) + \ln\left(\sum_{k=1}^{A^n} \frac{1}{k}\right) \\ &\leq \ln I_d(\boldsymbol{x}) + \ln(\ln A^n + 1) \\ &= \ln I_d(\boldsymbol{x}) + \ln(n \ln A + 1) \\ &\leq \ln I_d(\boldsymbol{x}) + \ln n + c, \end{aligned} \tag{23}$$

where $c = \ln(\ln A + 1)$. Therefore, the expected codeword length for $\boldsymbol{x}$ w.r.t. the randomness of the code

$$\begin{aligned} \boldsymbol{E}\{L_d(\boldsymbol{x})\} &\leq \boldsymbol{E}\{\ln I_d(\boldsymbol{x})\} + \ln n + c \\ &\leq \ln \boldsymbol{E}\{I_d(\boldsymbol{x})\} + \ln n + c \\ &= \ln\left(\sum_{k=1}^{A^n} k \cdot \left(1 - P_{\mathrm{s}}^d[\boldsymbol{x}]\right)^{k-1} \cdot P_{\mathrm{s}}[\boldsymbol{x}] + A^n \cdot (1 - P_{\mathrm{s}}[\boldsymbol{x}])^{A^n}\right) + \ln n + c \\ &= \ln\left(\sum_{k=1}^{\infty} \min\{k, A^n\} \cdot \left(1 - P_{\mathrm{s}}^d[\boldsymbol{x}]\right)^{k-1} \cdot P_{\mathrm{s}}^d[\boldsymbol{x}]\right) + \ln n + c \\ &\leq \ln\left\{\sum_{k=1}^{\infty} k \cdot \left(1 - P_{\mathrm{s}}^d[\boldsymbol{x}]\right)^{k-1} \cdot P_{\mathrm{s}}^d[\boldsymbol{x}]\right\} + \ln n + c \\ &= \ln\left(\frac{1}{P_{\mathrm{s}}^d[\boldsymbol{x}]}\right) + \ln n + c \\ &\leq nR_d(D, \hat{P}_{\boldsymbol{x}}) + \left(\frac{J}{2} + 1\right) \ln n + c', \end{aligned} \tag{24}$$

where c' is a constant, and where in the last step, we have used eq. (11).

Our goal, in this section, however, is more ambitious than that. We wish to prove the existence of a codebook with the following properties: (a) $L_d(\boldsymbol{x})$ is upper bounded in terms of $R_d(D, \hat{P}_{\boldsymbol{x}})$ plus some redundancy terms for every $\boldsymbol{x} \in \mathcal{X}^n$ and bounded d, and (b) The distortion constraint is met for every $\boldsymbol{x} \in \mathcal{X}^n$ and every distortion measure d with a given $d_{\max} < \infty$. To prove the second property, our approach is similar to that of Mahmood and Wagner [9]: We consider a fine grid, $\mathcal{D}_n$, in the space of distortion matrices, $\mathcal{D} = [0, d_{\max}]^{JK}$, where for each entry of the distortion matrix, there are n grid points with spacings of $d_{\max}/n$, that is $\mathcal{D}_n = \{0 \cdot d_{\max}/n, 1 \cdot d_{\max}/n, 2 \cdot d_{\max}/n, \ldots, n \cdot d_{\max}/n\}^{JK}$. If we can prove that there exists a codebook where property (b) holds just for every $d \in \mathcal{D}_n$, then for every $d \in \mathcal{D}$, the distortion cannot exceed $D + d_{\max}/n$. It should be pointed out that the choice of n as the number of grid points for each entry d is rather arbitrary, and can be viewed just as an example. In fact, one can afford even an exponentially fine resolution (and hence an exponentially decaying distortion redundancy), and our result will still hold. 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 [9].

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

Theorem 1 *Let $\epsilon > 0$ be arbitrarily small. For all sufficiently large n, there exists a codebook $\mathcal{C}_n = \{\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \ldots, \hat{\boldsymbol{x}}_{A^n}\}$, such that for every $\boldsymbol{x} \in \mathcal{X}^n$ and every $d \in \mathcal{D}_n$, the following two properties hold at the same time:*

(a) If $R_d(D, \hat{P}_{\boldsymbol{x}}) > 0$,

$$\frac{L_d(\boldsymbol{x})}{n} \le R_d(D, \hat{P}_{\boldsymbol{x}}) + \left(\frac{J}{2} + 2 + \epsilon\right) \cdot \frac{\ln n}{n} + o\left(\frac{\ln n}{n}\right). \tag{25}$$

If $R_d(D, \hat{P}_{\boldsymbol{x}}) = 0$,

$$\frac{L_d(\boldsymbol{x})}{n} \le (2 + \epsilon) \cdot \frac{\ln n}{n} + o\left(\frac{\ln n}{n}\right). \tag{26}$$

(b) $d(\boldsymbol{x}, \hat{\boldsymbol{x}}) \le nD$.

The main redundancy term in part (a), namely,

$$\left(\frac{J}{2} + 2 + \epsilon\right) \frac{\ln n}{n},$$

should be compared with those of Mahmood and Wagner [9], where the coefficients in front of $(\ln n)/n$ are, respectively, $2JK + J + 3$, $JK + J$, and $J^2K^2 + J - 2$, in Theorems 1, 2, and 3 of [9]. The differences are quite significant, especially for large J and K.

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 $R_d(D, \hat{P}_{\boldsymbol{x}}) > 0$, but the case $R_d(D, \hat{P}_{\boldsymbol{x}}) = 0$ can easily be handled in the very same manner.

Consider the quantity

$$E_n \stackrel{\Delta}{=} \boldsymbol{E}\left\{\max\left(\max_{d\in\mathcal{D}_n}\max_{\boldsymbol{x}\in\mathcal{X}^n}\mathcal{I}\{d(\boldsymbol{x},\hat{\boldsymbol{X}})>nD\},\right.\right.$$
$$\left.\left.\left[\max_{d\in\mathcal{D}_n}\max_{\boldsymbol{x}\in\mathcal{X}^n}\left(L_d(\boldsymbol{x})-nR_d(D;\hat{P}_{\boldsymbol{x}})-\left(\frac{J}{2}+2+\epsilon\right)\ln n-c\right)\right]_+\right)\right\}, \quad (27)$$

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

$$\max_{d\in\mathcal{D}_n}\max_{\boldsymbol{x}\in\mathcal{X}^n}\mathcal{I}\{d(\boldsymbol{x},\hat{\boldsymbol{x}})>nD\}\le\delta_n \quad (28)$$

and

$$\max_{d\in\mathcal{D}_n}\max_{\boldsymbol{x}\in\mathcal{X}^n}\left(L_d(\boldsymbol{x})-nR_d(D;\hat{P}_{\boldsymbol{x}})-\left(\frac{J}{2}+2+\epsilon\right)\log n-c\right)\le\delta_n \quad (29)$$

at the same time. Observe that since the left-hand side of (28) 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 $\boldsymbol{x}\in\mathcal{X}^n$ and $d\in\mathcal{D}_n$. Also, by (29), for the same codebook, we must have

$$L_d(\boldsymbol{x})\le nR_d(D;\hat{P}_{\boldsymbol{x}})+\left(\frac{J}{2}+2+\epsilon\right)\ln n+c+\delta_n \quad \forall\ \boldsymbol{x}\in\mathcal{X}^n,\ d\in\mathcal{D}_n, \quad (30)$$

where the extra term, δ_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

$$E_n \le \boldsymbol{E}\left\{\max_{d\in\mathcal{D}_n}\max_{\boldsymbol{x}\in\mathcal{X}^n}\mathcal{I}\{d(\boldsymbol{x},\hat{\boldsymbol{X}})>nD\}\right\}+$$
$$\boldsymbol{E}\left\{\left[\max_{d\in\mathcal{D}_n}\max_{\boldsymbol{x}\in\mathcal{X}^n}\left(L_d(\boldsymbol{x})-nR_d(D,\hat{P}_{\boldsymbol{x}})-\left(\frac{J}{2}+2+\epsilon\right)\ln n-c\right)\right]_+\right\}, \quad (31)$$

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

$$\begin{aligned}
&\boldsymbol{E}\left\{\max_{d\in\mathcal{D}_n}\max_{\boldsymbol{x}\in\mathcal{X}^n}\mathcal{I}\{d(\boldsymbol{x},\hat{\boldsymbol{X}})>nD\}\right\}\\
\le\ &\boldsymbol{E}\left\{\sum_{d\in\mathcal{D}_n}\sum_{\boldsymbol{x}\in\mathcal{X}^n}\mathcal{I}\{d(\boldsymbol{x},\hat{\boldsymbol{X}})>nD\}\right\}\\
=\ &\sum_{d\in\mathcal{D}_n}\sum_{\boldsymbol{x}\in\mathcal{X}^n}\boldsymbol{E}\left\{\mathcal{I}\{d(\boldsymbol{x},\hat{\boldsymbol{X}})>nD\}\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{d\in\mathcal{D}_n}\sum_{\boldsymbol{x}\in\mathcal{X}^n} \Pr\{d(\boldsymbol{x},\hat{\boldsymbol{X}}) > nD\} \\
&= \sum_{d\in\mathcal{D}_n}\sum_{\boldsymbol{x}\in\mathcal{X}^n} \left(1-P_{\mathrm{s}}^d[\boldsymbol{x}]\right)^{A^n} \\
&\le \sum_{d\in\mathcal{D}_n}\sum_{\boldsymbol{x}\in\mathcal{X}^n} \exp\left\{-A^n P_{\mathrm{s}}^d[\boldsymbol{x}]\right\} \\
&\le \sum_{d\in\mathcal{D}_n}\sum_{\boldsymbol{x}\in\mathcal{X}^n} \exp\left\{-\exp\left\{n\left[\ln A - R_d(D,\hat{P}_{\boldsymbol{x}}) - O\left(\frac{\ln n}{n}\right)\right]\right\}\right\} \\
&\le (n+1)^{JK}\cdot J^n \exp\left(-\exp\left\{n\left[\ln A - \ln J - O\left(\frac{\ln n}{n}\right)\right]\right\}\right), \qquad (32)
\end{aligned}
$$

which indeed decays as $n\to\infty$, since we have assumed that $A > J$. As for the second term of (31), we have:

$$
\begin{aligned}
&\boldsymbol{E}\left\{\left[\max_{d\in\mathcal{D}_n}\max_{\boldsymbol{x}\in\mathcal{X}^n}\left(L_d(\boldsymbol{x}) - nR_d(D,\hat{P}_{\boldsymbol{x}}) - \left(\frac{J}{2}+2+\epsilon\right)\ln n - c\right)\right]_+\right\} \\
&= \boldsymbol{E}\left\{\left[\max_{d\in\mathcal{D}_n}\max_{\boldsymbol{x}\in\mathcal{X}^n}\left(\ln I_d(\boldsymbol{x}) - nR_d(D,\hat{P}_{\boldsymbol{x}}) - \left(\frac{J}{2}+1+\epsilon\right)\ln n\right)\right]_+\right\} \\
&= \int_0^\infty \Pr\left\{\max_{d\in\mathcal{D}_n}\max_{\boldsymbol{x}\in\mathcal{X}^n}\left[\ln I_d(\boldsymbol{x}) - nR_d(D,\hat{P}_{\boldsymbol{x}}) - \left(\frac{J}{2}+1+\epsilon\right)\ln n\right] \ge s\right\}\mathrm{d}s \\
&= \int_0^{n\ln A} \Pr\left\{\max_{d\in\mathcal{D}_n}\max_{\boldsymbol{x}\in\mathcal{X}^n}\left[\ln I_d(\boldsymbol{x}) - nR_d(D,\hat{P}_{\boldsymbol{x}}) - \left(\frac{J}{2}+1+\epsilon\right)\ln n\right] \ge s\right\}\mathrm{d}s \\
&\le \int_0^{n\ln A} \Pr\bigcup_{d\in\mathcal{D}_n}\bigcup_{\boldsymbol{x}\in\mathcal{X}^n}\left\{I_d(\boldsymbol{x}) \ge \exp\left[nR_d(D,\hat{P}_{\boldsymbol{x}}) + \left(\frac{J}{2}+1+\epsilon\right)\ln n + s\right]\right\}\mathrm{d}s \\
&\le \sum_{d\in\mathcal{D}_n}\sum_{\boldsymbol{x}\in\mathcal{X}^n}\int_0^{n\ln A} \Pr\left\{I_d(\boldsymbol{x}) \ge \exp\left[nR_d(D,\hat{P}_{\boldsymbol{x}}) + \left(\frac{J}{2}+1+\epsilon\right)\ln n + s\right]\right\}\mathrm{d}s \\
&\le \sum_{d\in\mathcal{D}_n}\sum_{\boldsymbol{x}\in\mathcal{X}^n}\int_0^{n\ln A} \Pr\left\{I_d(\boldsymbol{x}) \ge \exp\left[nR_d(D,\hat{P}_{\boldsymbol{x}}) + \left(\frac{J}{2}+1+\epsilon\right)\ln n\right]\right\}\mathrm{d}s \\
&\le (n\ln A)\cdot\sum_{d\in\mathcal{D}_n}\sum_{\boldsymbol{x}\in\mathcal{X}^n}(1-P_{\mathrm{s}}^d[\boldsymbol{x}])^{\exp\left[nR_d(D,\hat{P}_{\boldsymbol{x}})+\left(\frac{J}{2}+1+\epsilon\right)\ln n\right]} \\
&= (n\ln A)\cdot\sum_{d\in\mathcal{D}_n}\sum_{\boldsymbol{x}\in\mathcal{X}^n}\exp\left\{\exp\left[nR_d(D,\hat{P}_{\boldsymbol{x}}) + \left(\frac{J}{2}+1+\epsilon\right)\ln n\right]\ln(1-P_{\mathrm{s}}^d[\boldsymbol{x}])\right\} \\
&\le (n\ln A)\cdot\sum_{d\in\mathcal{D}_n}\sum_{\boldsymbol{x}\in\mathcal{X}^n}\exp\left\{-\exp\left[nR_d(D,\hat{P}_{\boldsymbol{x}}) + \left(\frac{J}{2}+1+\epsilon\right)\ln n\right]P_{\mathrm{s}}^d[\boldsymbol{x}]\right\} \\
&\le (n\ln A)\cdot\sum_{d\in\mathcal{D}_n}\sum_{\boldsymbol{x}\in\mathcal{X}^n}\exp\left\{-\exp\left[(1+\epsilon)\ln n\right]\right\} \\
&= (n\ln A)\cdot(n+1)^{JK}\cdot J^n\cdot\exp\{-n^{1+\epsilon}\}, \qquad (33)
\end{aligned}
$$

which decays as well. This completes the proof of Theorem 1.

5 Beyond Memoryless Sources and Additive Distortion Measures

Our results in Sections 3 and 4 hold *pointwise*, for each and every individual source vector $\boldsymbol{x}$, even without taking the expectation w.r.t. the randomness of the source vector. Of course, one can also take the expectation and obtain a result on the rate redundancy relative to the expectation of the empirical rate-distortion function, $R_d(D, \hat{P}_{\boldsymbol{x}})$ (which in turn converges almost surely to $R_d(D, P)$), as was actually done in [10, Theorems 1–4]. But in spite of the pointwise nature of our results so far, the codes that we have been considering are suitable only for the class of memoryless sources and additive distortion measures, since the length function, $L_d(\boldsymbol{x})$, whose main term is $nR_d(D, \hat{P}_{\boldsymbol{x}})$, depends on $\boldsymbol{x}$ only via its zeroth order empirical distribution, which is blind to any empirical dependencies and repetitive patterns within the source sequence, $\boldsymbol{x}$.

In this section, we would like to remain in the realm of individual sequences, but to expand the scope to codes that are suitable beyond memoryless sources, i.e., codes that are designed to exploit the memory within the given source sequence to be compressed. By the same token, we will be interested in more general classes of distortion measures, not necessarily additive ones. In this section, the discussion will be less formal than before, as we will only outline how the ideas of the previous sections extend to this more general setting, without any heavy analysis of exact redundancy rates.

We adopt the individual-sequence setting, in the footsteps of Lempel and Ziv [22]. According to this setting, defined in [22] for the lossless case, the source sequence, $\boldsymbol{x}$, is a given deterministic setting, but the encoder is limited to be implementable by an information lossless finite-state machine with s states, and the asymptotic regime is that $s \ll n$, as the limit $s \to \infty$ is taken after the limit $n \to \infty$.

When it comes to source coding with distortion, a natural extension of this setting could be based on the fact that in lossy compression, there is no loss of optimality if the encoder is implemented as a cascade of two mappings, as follows: first, apply a reproduction encoder (or, vector quantizer), that maps the source $\boldsymbol{x}$ directly to its reproduction, $\hat{\boldsymbol{x}}$, and then compress $\hat{\boldsymbol{x}}$ by a lossless encoder, without any additional distortion. Accordingly, we can adopt this structure with the limitation that the lossless encoder of the second stage is a finite-state encoder with s states, exactly as in

[22].[4] Applying, the converse theorem of Lempel and Ziv [22, Theorem 1], we have that the length of the lossless code associated with the reproduction vector, $\hat{\boldsymbol{x}}$, is lower bounded by

$$L(\hat{\boldsymbol{x}}) \geq [c(\hat{\boldsymbol{x}}) + s^2] \log \frac{c(\hat{\boldsymbol{x}}) + s^2}{4s^2} + 2s^2, \tag{34}$$

where $c(\hat{\boldsymbol{x}})$ is the largest number of distinct phrases whose concatenation forms $\hat{\boldsymbol{x}}$. Since $\hat{\boldsymbol{x}}$ is constrained to lie within distance nD away from $\boldsymbol{x}$, we reach at the obvious lower bound of

$$L_d(\boldsymbol{x}) \geq \min_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}}) \leq nD\}} \left\{ [c(\hat{\boldsymbol{x}}) + s^2] \log \frac{c(\hat{\boldsymbol{x}}) + s^2}{4s^2} + 2s^2 \right\}, \tag{35}$$

and a conceptually simple way to asymptotically achieve this lower bound is to choose, among all vectors, $\{\hat{\boldsymbol{x}}\}$, within distortion nD away from $\boldsymbol{x}$, the one whose Lempel-Ziv (LZ) code-length is minimal, and to transmit its compressed from using the LZ algorithm [22, Theorem 2]. The LZ codelength of $\hat{\boldsymbol{x}}$, which we denote by $LZ(\hat{\boldsymbol{x}})$, is upper bounded by

$$LZ(\hat{\boldsymbol{x}}) \leq [c_{\text{LZ}}(\hat{\boldsymbol{x}}) + 1] \log(2K[c_{\text{LZ}}(\hat{\boldsymbol{x}}) + 1]), \tag{36}$$

where $c_{\text{LZ}}(\hat{\boldsymbol{x}})$ is the number of phrases of $\hat{\boldsymbol{x}}$ obtained by the incremental parsing procedure of the LZ algorithm [22, proof of Theorem 2]. Note that here, $d(\boldsymbol{x}, \hat{\boldsymbol{x}})$ can be any distortion function, not necessarily an additive one. The painful part of this achievability scheme, however, is the exponential complexity associated with the search across the 'sphere', $\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x}, \hat{\boldsymbol{x}}) \leq nD\}$. In the case of an additive distortion measure, the complexity of this search grows at the exponential rate of $\exp\{nE(D)\}$, where $E(D) = \max H(\hat{X}|X)$, with X being a dummy random variable, distributed according to $\hat{P}_{\boldsymbol{x}}$, and with the maximization being taken over all conditional distributions, $\{P_{\hat{X}|X}\}$, such that $\boldsymbol{E}\{d(X, \hat{X})\} \leq D$. When D is relatively large, then so is $E(D)$.

We now propose an alternative approach to this problem using the ideas of the previous section. To this end, we first have to extend the random coding distribution, W, to be suitable beyond the class of memoryless sources. Following the findings of [5] and [13], consider the random coding distribution,

$$W(\hat{\boldsymbol{x}}) = \frac{2^{-LZ(\hat{\boldsymbol{x}})}}{\sum_{\hat{\boldsymbol{x}}' \in \hat{\mathcal{X}}^n} 2^{-LZ(\hat{\boldsymbol{x}}')}}. \tag{37}$$

The associated single success probability is given by

$$P_{\text{s}}^d[\boldsymbol{x}] = \sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}}) \leq nD\}} W(\hat{\boldsymbol{x}}). \tag{38}$$

[4]Note that this setting is somewhat different form Ziv's model of lossy compression for individual sequences, [21].

We can repeat the same derivations as in Section 4, but with the new expression of $P_{\rm s}^d[\boldsymbol{x}]$, and use eqs. (20) and (21) to argue that we can achieve compression according to the length function,

$$L_d(\boldsymbol{x}) = -\log P_{\rm s}^d[\boldsymbol{x}] + (2+\epsilon)\log n \tag{39}$$

within distortion nD (w.r.t. any distortion measure d within a class $\mathcal{D}$ of distortion measures that can be well approximated using a grid whose size is no more than exponential), pointwise, for every $\boldsymbol{x}$, similarly as before. Now, observe that the main term of $L_d(\boldsymbol{x})$, namely, $-\log P_{\rm s}^d[\boldsymbol{x}]$, can be upper bounded as follows.

$$\begin{aligned} -\log P_{\rm s}^d[\boldsymbol{x}] &= -\log\left[\sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}} \frac{2^{-LZ(\hat{\boldsymbol{x}})}}{\sum_{\hat{\boldsymbol{x}}'} 2^{-LZ(\hat{\boldsymbol{x}}')}}\right] \\ &\le -\log\left[\sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}} 2^{-LZ(\hat{\boldsymbol{x}})}\right] \\ &\le -\log\left[\max_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}} 2^{-LZ(\hat{\boldsymbol{x}})}\right] \\ &= \min_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}} LZ(\hat{\boldsymbol{x}}), \end{aligned} \tag{40}$$

where in the second line, we used Kraft's inequality. This means that this scheme also asymptotically achieves the lower bound (35). However, this coding scheme has a different computational complexity than the earlier one. The number of metric calculations that this encoder has to carry out until it finds the first codeword within distortion nD, is a random variable, but it is typically of the order of magnitude of $1/P_{\rm s}^d[\boldsymbol{x}]$. Which one of the encoders is better in terms of the computational complexity? The answer depends, of course, on $\boldsymbol{x}$ and D. For small D, it is more efficient to use the first approach, as $e^{nE(D)}$ is relatively small, whereas $1/P_{\rm s}^d[\boldsymbol{x}]$ is relatively large. On the other hand, for large D, the contrary is true. In fact, by Ziv's inequality [3, p. 455, eq. (13.125)] (applied to memoryless sources), it is readily seen that $1/P_{\rm s}^d[\boldsymbol{x}] \stackrel{\cdot}{\le} \exp\{nR(D,\hat{P}_{\boldsymbol{x}})\}$, and so, whenever $R(D,\hat{P}_{\boldsymbol{x}}) < E(D)$, it is definitely better to use the second scheme.

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

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

[5] A. Cohen and N. Merhav, "Universal randomized guessing subjected to distortion," submitted to *IEEE Trans. Inform. Theory*, December 2021. Available on-line at: `https://arxiv.org/pdf/2112.13594.pdf`

[6] N. G. de Bruijn, *Asymptotic Methods in Analysis*, Dover Publications, Inc., New York 1981.

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

[8] 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.

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

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

[11] 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.

[12] N. Merhav, "Statistical physics and information theory," *Foundations and Trends in Communications and Information Theory*, vol. 6, nos. 1–2, pp. 1–212, 2009.

[13] 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.

[14] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2124–2147, October 1998.

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

[16] J. F. Silva and P. Piantanida, "On universal d-semifaithful coding for memoryless sources with infinite alphabets," `https://arxiv.org/pdf/2107.05082.pdf`

[17] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

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

[19] 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. Inform. Theory*, vol. 47, no. 1, pp. 126-145, January 2001.

[20] 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.

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

[22] 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.